BI

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/10, 1/19, C12Q 1/68

A1

(11) International Publication Number: WO 94/10300

(43) International Publication Date: 11 May 1994 (11.05.94)

(21) International Application Number: PCT/US93/10069

(22) International Filing Date: 20 October 1993 (20.10.93)

(30) Priority data:
07/969,038 30 October 1992 (30.10.92) US

(71) Applicant: THE GENERAL HOSPITAL CORPORATION [US/US]; 55 Fruit Street, Boston, MA 02114 (US).

(72) Inventors: BRENT, Roger ; 27 Wendell Street, Cambridge, MA 02138 (US). GYURIS, Jeno ; 43 Hemlock Street, Arlington, MA 02174 (US). GOLEMIS, Erica ; 29 Quincy Street, Somerville, MA 02143 (US).

(74) Agent: CLARK, Paul, T.; Fish & Richardson, 225 Franklin Street, Boston, MA 02110-2804 (US).

(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: INTERACTION TRAP SYSTEM FOR ISOLATING NOVEL PROTEINS

(57) Abstract

Disclosed is a method for determining whether a first protein is capable of physically interacting with a second protein. The method involves: (a) providing a host cell which contains (i) a reporter gene operably linked to a protein binding site; (ii) a first fusion gene which expresses a first fusion protein, the first fusion protein including the first protein covalently bonded to a binding moiety which is capable of specifically binding to the protein binding site; and (iii) a second fusion gene which expresses a second fusion protein, the second fusion protein including the second protein covalently bonded to a weak gene activating moiety; and (b) measuring expression of the reporter gene as a measure of an interaction between the first and the second proteins. Such a determination facilitates the isolation of the gene encoding the interacting protein. Also disclosed herein is recombinant Cdi1 polypeptide, nucleic acid encoding the Cdi1 polypeptide, and uses thereof.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

INTERACTION TRAP SYSTEM FOR ISOLATING NOVEL PROTEINS

Background of the Invention

This invention was made with Government support awarded by the National Institute of Health. The government has certain rights in the invention. This invention relates to methods for isolating novel proteins. This invention also relates to cancer diagnostics and therapeutics.

In most eukaryotic cells, the cell cycle is governed by controls exerted during G1 and G2. During G2, cells decide whether to enter M in response to relatively uncharacterized intracellular signals, such as those that indicate completion of DNA synthesis (Nurse, Nature 344:503-508, 1990; Enoch and Nurse, Cell 65:921-923, 1991). During G1, cells either enter S or withdraw from the cell cycle and enter a nondividing state known as G0 (Pardee, Science 246:603-608, 1989). While the control mechanisms for these decisions are not yet well understood, their function is clearly central to processes of normal metazoa development and to carcinogenesis.

In yeast, and probably in all eukaryotes, the G1/S and G2/M transitions depend on a family of ~34kd protein kinases, the Cdc2 proteins, encoded by the *cdc2*$^+$ (in *S. pombe*) and *CDC28* (in *S. cerevisiae*) genes. Cdc2 family proteins from mammalian cells have been also identified. Some including Cdc2 (Lee and Nurse, Nature 327:31-35, 1987), Cdk2 (Elledge and Spotswood, EMBO J. 10:2653-2659, 1991; Tsai et al., Nature 353:174-177, 1991), and Cdk3 (Meyerson et al., EMBO J. 11:2909-2917, 1992) can complement a *cdc28*$^-$ *S. cerevisiae* for growth.

The activity of the Cdc2 proteins at the G2/M transition point is regulated in two ways: positively, by association with regulatory proteins called cyclins, and

negatively, by phosphorylation of a tyrosine near their ATP binding site. At least one of these regulatory mechanisms is operative during G1 (see Figure 1A). At this time, Cdc2 protein activity is regulated by facultative association with different G1 specific cyclins. In *S. cerevisiae* at least five putative G1 cyclins have been identified in genetic screens, including the products of the *CLN1, CLN2, CLN3, HSC26* and *CLB5* genes (Cross, Mol. Cell. Biol 8:4675-4684, 1988; Nash et al., EMBO J. 7:4335-4346, 1988; Hadwiger et al., Proc. Nat. Acad. Sci. USA 86:6255-6259, 1989; and Ogas et al., Cell 66:1015-1026, 1991). The *CLN1, CLN2,* and *CLN3* proteins (here called Cln1, Cln2, and Cln3) are each individually sufficient to permit a cell to make the G1 to S transition (Richardson et al., Cell 59:1127-1133, 1989), and at least one of them (Cln2) associates with Cdc28 in a complex that is active as a protein kinase (Wittenberg et al., Cell 62:225-237, 1990). Recently, putative G1 cyclins have been identified in mammalian cells: Cyclin C, Cyclin D (three forms), and Cyclin E (Koff et al., Cell 66:1217-1228, 1991; Xiong et al., Cell 65:691-699, 1991). Each of these three mammalian cyclins complement a yeast deficient in Cln1, Cln2, and Cln3, and each is expressed during G1.

In *S. cerevisiae*, the synthesis, and in some cases, the activity of the G1 cyclins is under the control of a network of genes that help to couple changes in the extracellular environment to G1 regulatory decisions (Figure 1A). For example, the *SWI4* and *SWI6* gene products positively regulate *CLN1* and *CLN2* transcription and may also positively modulate the activity of Cln3 (Nasmyth and Dirick, Cell 66:995-1013, 1991), the *FAR1* product negatively regulates both *CLN2* transcription and the activity of its product (Chang and Herskowitz, Cell 63:999-1011, 1990), and the *FUS3* product

negatively regulates Cln3 activity (Elion et al., Cell 60:649-664, 1990).

Several lines of evidence suggest that mammalian G1 to S transitions may be regulated by similar mechanisms: regulatory molecules (Cdc2 kinases and cyclins) similar to those found in yeast are observed in mammalian G1, and like *S. cerevisiae*, mammalian cells arrest in G1 when deprived of nutrients and in response to certain negative regulatory signals, including contact with other cells or treatment with negative growth factors (e.g., TGF-β) (Figure 1B). However, several considerations suggest that the higher eukaryotic G1 regulatory machinery is likely to be more sophisticated than that of yeast. First, in mammalian cells there appear to be more proteins involved in the process. At least ten different Cdc2 family proteins and related protein kinases (see Meyerson et al., EMBO J. 11:2909-2917, 1992) and at least three distinct classes of putative G1 cyclins (Koff et al., Cell 66:1217-1228, 1991; Matsushime et al., Cell 65:701-713, 1991; Motokura et al., Nature 339:512-518, 1991; Xiong et al., Cell 65:691-699, 1991) have been identified. Second, unlike yeast, the proliferation of most mammalian cells depends on extracellular protein factors (in particular, positive growth regulatory proteins), deprivation of which leads to arrest in G1. Third, arrest of many cell types during G1 can progress to a state, G0, that may not strictly parallel any phase of the yeast cell cycle.

Because proteins involved in controlling normal cell division decisions in mammals (e.g., humans) are also very likely to play a key role in malignant cell growth, identification and isolation of such proteins facilitate the development of useful cancer diagnostics as well as anti-cancer therapeutics. We now describe (i) a novel system for the identification of proteins which,

at some time during their existence, participate in a particular protein-protein interaction; (ii) the use of this system to identify interacting proteins which are key regulators of mammalian cell division; and (iii) one such interacting protein, termed Cdi1, a cell cycle control protein which provides a useful tool for cancer diagnosis and treatment.

Summary of the Invention

In general, the invention features a method for determining whether a first protein is capable of physically interacting (i.e., directly or indirectly) with a second protein. The method involves: (a) providing a host cell which contains (i) a reporter gene operably linked to a protein binding site; (ii) a first fusion gene which expresses a first fusion protein, the first fusion protein including the first protein covalently bonded to a binding moiety which is capable of specifically binding to the protein binding site; and (iii) a second fusion gene which expresses a second fusion protein, the second fusion protein including the second protein covalently bonded to a weak gene activating moiety; and (b) measuring expression of the reporter gene as a measure of an interaction between the first and the second proteins. In a preferred embodiment, the method further involves isolating the gene encoding the second protein.

In other preferred embodiments, the weak gene activating moiety is of lesser activation potential than GAL4 activation region II and preferably is the gene activating moiety of B42 or a gene activating moiety of lesser activation potential; the host cell is a yeast cell; the reporter gene includes the *LEU2* gene or the *lacZ* gene; the host cell further contains a second reporter gene operably linked to the protein binding site, for example, the host cell includes both a *LEU2*

reporter gene and a *lacZ* reporter gene; the protein binding site is a LexA binding site and the binding moiety includes a LexA DNA binding domain; the second protein is a protein involved in the control of eukaryotic cell division, for example, a Cdc2 cell division control protein.

In a second aspect, the invention features a substantially pure preparation of Cdi1 polypeptide. Preferably, the Cdi1 polypeptide includes an amino acid sequence substantially identical to the amino acid sequence shown in Figure 6 (SEQ ID NO: 1); and is derived from a mammal, for example, a human.

In a related aspect, the invention features purified DNA (for example, cDNA) which includes a sequence encoding a Cdi1 polypeptide, and preferably a human Cdi1 polypeptide, of the invention.

In other related aspects, the invention features a vector and a cell which includes a purified DNA of the invention; a purified antibody which specifically binds a Cdi1 polypeptide of the invention; and a method of producing a recombinant Cdi1 polypeptide invloving, providing a cell transformed with DNA encoding a Cdi1 polypeptide positioned for expression in the cell; culturing the transformed cell under conditions for expressing the DNA; and isolating the recombinant Cdi1 polypeptide. The invention further features recombinant Cdi1 polypeptide produced by such expression of a purified DNA of the invention.

In yet another aspect, the invention features a therapeutic composition which includes as an active ingredient a Cdi1 polypeptide of the invention, the active ingredient being formulated in a physiologically-acceptable carrier. Such a therapeutic composition is useful in a method of inhibiting cell proliferation in a mammal, involving administering the therapeutic

composition to the mammal in a dosage effective to inhibit mammalian cell division.

In a final aspect, the invention features a method of detecting a malignant cell in a biological sample, involving measuring Cdi1 gene expression in the sample, a change in Cdi1 expression relative to a wild-type sample being indicative of the presence of the malignant cell.

As used herein, by "reporter gene" is meant a gene whose expression may be assayed; such genes include, without limitation, *lacZ*, amino acid biosynthetic genes, e.g. the yeast *LEU2*, *HIS3*, *LYS2*, or *URA3* genes, nucleic acid biosynthetic genes, the mammalian chloramphenicol transacetylase (CAT) gene, or any surface antigen gene for which specific antibodies are available.

By "operably linked" is meant that a gene and a regulatory sequence(s) are connected in such a way as to permit gene expression when the appropriate molecules (e.g., transcriptional activator proteins or proteins which include transcriptional activation domains) are bound to the regulatory sequence(s).

By a "binding moiety" is meant a stretch of amino acids which is capable of directing specific polypeptide binding to a particular DNA sequence (i.e., a "protein binding site").

By "weak gene activating moiety" is meant a stretch of amino acids which is capable of weakly inducing the expression of a gene to whose control region it is bound. As used herein, "weakly" is meant below the level of activation effected by GAL4 activation region II (Ma and Ptashne, Cell 48:847, 1987) and is preferably at or below the level of activation effected by the B42 activation domain of Ma and Ptashne (Cell 51:113, 1987). Levels of activation may be measured using any downstream reporter gene system and comparing, in parallel assays, the level of expression stimulated by the GAL4 region II-

polypeptide with the level of expression stimulated by the polypeptide to be tested.

By "substantially pure" is meant a preparation which is at least 60% by weight (dry weight) the compound of interest, e.g., a Cdi1 polypeptide. Preferably the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight the compound of interest. Purity can be measured by any appropriate method, e.g., column chromatography, polyacrylamide gel electrophoresis, or HPLC analysis.

By "purified DNA" is meant DNA that is not immediately contiguous with both of the coding sequences with which it is immediately contiguous (one on the 5′ end and one on the 3′ end) in the naturally occurring genome of the organism from which it is derived. The term therefore includes, for example, a recombinant DNA which is incorporated into a vector; into an autonomously replicating plasmid or virus; or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate molecule (e.g., a cDNA or a genomic DNA fragment produced by PCR or restriction endonuclease treatment) independent of other sequences. It also includes a recombinant DNA which is part of a hybrid gene encoding additional polypeptide sequence.

By "substantially identical" is meant an amino acid sequence which differs only by conservative amino acid substitutions, for example, substitution of one amino acid for another of the same class (e.g., valine for glycine, arginine for lysine, etc.) or by one or more non-conservative substitutions, deletions, or insertions located at positions of the amino acid sequence which do not destroy the function of the protein (assayed, e.g., as described herein). A "substantially identical" nucleic acid sequence codes for a substantially identical amino acid sequence as defined above.

By "transformed cell" is meant a cell into which (or into an ancestor of which) has been introduced, by means of recombinant DNA techniques, a DNA molecule encoding (as used herein) a Cdi1 polypeptide.

By "positioned for expression" is meant that the DNA molecule is positioned adjacent to a DNA sequence which directs transcription and translation of the sequence (i.e., facilitates the production of, e.g., a Cdi1 polypeptide).

By "purified antibody" is meant antibody which is at least 60%, by weight, free from the proteins and naturally-occurring organic molecules with which it is naturally associated. Preferably, the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight, antibody, e.g., Cdi1-specific antibody. A purified Cdi1 antibody may be obtained, for example, by affinity chromatography using recombinantly-produced Cdi1 polypeptide and standard techniques.

By "specifically binds" is meant an antibody which recognizes and binds Cdi1 polypeptide but which does not substantially recognize and bind other molecules in a sample, e.g., a biological sample, which naturally includes Cdi1 polypeptide.

By a "malignant cell" is meant a cell which has been released from normal cell division control. Included in this definition are transformed and immortalized cells.

The interaction trap system described herein provides advantages over more conventional methods for isolating interacting proteins or genes encoding interacting proteins. Most notably, applicants' system provides a rapid and inexpensive method having very general utility for identifying and purifying genes

encoding a wide range of useful proteins based on the protein's physical interaction with a polypeptide of known diagnostic or therapeutic usefulness. This general utility derives in part from the fact that the components of the system can be readily modified to facilitate detection of protein interactions of widely varying affinity (e.g., by using reporter genes which differ quantitatively in their sensitivity to a protein interaction). The inducible nature of the promoter used to express the interacting proteins also increases the scope of candidate interactors which may be detected since even proteins whose chronic expression is toxic to the host cell may be isolated simply by inducing a short burst of the protein's expression and testing for its ability to interact and stimulate expression of a β-galactosidase reporter gene.

Moreover, detection of interacting proteins through the use of a weak gene activation domain tag avoids the restrictions on the pool of available candidate interacting proteins which is characteristically associated with stronger activation domains (such as GAL4 or VP16); although the mechanism is unclear, such a restriction apparently results from low to moderate levels of host cell toxicity mediated by the strong activation domain.

Other features and advantages of the invention will be apparent from the following detailed description thereof, and from the claims.

Brief Description of the Drawings

The drawings are first briefly described.

FIGURE 1 illustrates cell cycle control systems. FIGURE 1(A) illustrates G1 control in yeast. FIGURE 1B illustrates cell cycle control in yeast and mammals.

FIGURE 2 A-C illustrates an interaction trap system according to the invention.

FIGURE 3A is a diagrammatic representation of a "bait" protein useful in the invention; the numbers represent amino acids. FIGURE 3B is a diagrammatic representation of reporter genes useful in the invention. FIGURE 3C is a diagrammatic representation of a library expression plasmid useful in the invention and the N-terminal amino acid sequence of an exemplary "prey" protein according to the invention.

FIGURE 4 depicts yeast assays demonstrating the specificity of the Cdi1/Cdc2 interaction.

FIGURE 5 shows the results of an immunoprecipitation experiment demonstrating that Cdi1 physically interacts with Cdc2.

FIGURE 6 shows the Cdi1 coding sequence together with the predicted amino-acid sequence of its open reading frame (SEQ ID NO:1).

In FIGURE 7A, the growth rates of yeast cells that express Cdi1 are depicted; open squares are cells transformed with expression vectors only; ovals are cells expressing Cdc2; triangles are cells expressing Cdi1; and filled squares are cells expressing Cdi1 and Cdc2. In FIGURE 7B is shown a budding index of yeast that express Cdi1. In FIGURE 7C is shown a FACS analysis of yeast that express Cdi1; fluorescence (on the x-axis) is shown as a function of cell number (on the Y-axis).

FIGURE 8A shows the morphology of control cells; FIGURE 8B shows the morphology of control cells stained with DAPI; FIGURE 8C shows the morphology of cells expressing Cdi1; and FIGURE 8D shows the morphology of cells expressing Cdi1 stained with DAPI.

FIGURE 9A indicates the timing of Cdi1 expression in Hela cells; lanes represent different timepoints: (1) 0h, (2) 3h, (3) 6h, (4) 9h, (5) 12h, (6) 15h, (7) 18h, (8) 21h, (9) 24h, and (10) 27h after release. FIGURE 9B shows the effect of Cdi1 overexpression.

FIGURE 10 shows an alignment of Cdc2 proteins and FUS3. Depicted is an alignment of the sequences of the bait proteins used herein. Amino acids are numbered as in human Cdc2. Abbreviations are as follows: HsCdc2, human Cdc2; HsCdk2, human Cdk2; ScCdc28, *S. cerevisiae* Cdc28; DmCdc2 and DmCdc2c, the two Drosophila Cdc2 isolates; and ScFus3, *S. cerevisiae* FUS3. Residues shown in boldface are conserved between the Cdc2 family members; residues present in Fus3 are also shown in bold. Asterisks indicate potential Cdi1 contact points, i.e., amino acids that are conserved among human Cdc2, Cdk2, *S. cerevisiae* Cdc28, and Drosophila Cdc2, but that differ in Drosophila Cdc2c and in Fus3.

There now follows a description of one example of an interaction trap system and its use for isolating a particular cell division protein. This example is designed to illustrate, not limit, the invention.

Detailed Description

Applicants have developed an in vivo interaction trap system for the isolation of genes encoding proteins which physically interact with a second protein of known diagnostic or therapeutic utility. The system involves a eukaryotic host strain (e.g., a yeast strain) which is engineered to express the protein of therapeutic or diagnostic interest as a fusion protein covalently bonded to a known DNA binding domain; this protein is referred to as a "bait" protein because its purpose in the system is to "catch" useful, but as yet unknown or uncharacterized, interacting polypeptides (termed the "prey"; see below). The eukaryotic host strain also contains one or more "reporter genes", i.e., genes whose transcription is detected in response to a bait-prey interaction. Bait proteins, via their DNA binding domain, bind to their specific DNA site upstream of a

reporter gene; reporter transcription is not stimulated, however, because the bait protein lacks its own activation domain.

To isolate genes encoding novel interacting proteins, cells of this strain (containing a reporter gene and expressing a bait protein) are transformed with individual members of a DNA (e.g., a cDNA) expression library; each member of the library directs the synthesis of a candidate interacting protein fused to a weak and invariant gene activation domain tag. Those library-encoded proteins that physically interact with the promoter-bound bait protein are referred to as "prey" proteins. Such bound prey proteins (via their activation domain tag) detectably activate the transcription of the downstream reporter gene and provide a ready assay for identifying particular cells which harbor a DNA clone encoding an interacting protein of interest.

One example of such an interaction trap system is shown in Figure 2. Figure 2A shows a yeast strain containing two reporter genes, *LexAop-LEU2* and *LexAop-lacZ*, and a constitutively expressed bait protein, LexA-Cdc2. Synthesis of prey proteins is induced by growing the yeast in the presence of galactose. Figure 2B shows that if the prey protein does not interact with the transcriptionally-inert LexA-fusion bait protein, the reporter genes are not transcribed; the cell cannot grow into a colony on leu^- medium, and it is white on Xgal medium because it contains no β-galactosidase activity. Figure 2C shows that, if the prey protein interacts with the bait, then both reporter genes are active; the cell forms a colony on leu^- medium, and cells in that colony have β-galactosidase activity and are blue on Xgal medium.

As described herein, in developing the interaction trap system shown diagrammatically in Figure 2, careful

attention was paid to three classes of components: (i) use of bait proteins that contained a site-specific DNA binding domain that was known to be transcriptionally inert; (ii) use of reporter genes that had essentially no basal transcription and that were bound by the bait protein; and (iii) use of library-encoded prey proteins, all of which were expressed as chimeras whose amino termini contained the same weak activation domain and, preferably, other useful moieties, such as nuclear localization signals.

Each component of the system is now described in more detail.

Bait Proteins

The selection host strain depicted in Figure 2 contains a Cdc2 bait and a DNA binding moiety derived from the bacterial LexA protein (see Figure 3A). The use of a LexA DNA binding domain provides certain advantages. For example, in yeast, the LexA moiety contains no activation function and has no known effect on transcription of yeast genes (Brent and Ptashne, Nature 312:612-615, 1984; Brent and Ptashne, Cell 43:729-736, 1985). In addition, use of the LexA rather than the GAL4 DNA-binding domain allows conditional expression of prey proteins in response to galactose induction; this facilitates detection of prey proteins which might be toxic to the host cell if expressed continuously. Finally, the use of LexA allows knowledge regarding the interaction between LexA and the LexA binding site (i.e., the LexA operator) to be exploited for the purpose of optimizing operator occupancy.

The bait protein illustrated in Figure 3A also includes a LexA dimerization domain; this optional domain facilitates efficient LexA dimer formation. Because LexA binds its DNA binding site as a dimer, inclusion of this domain in the bait protein also optimizes the efficiency

of operator occupancy (Golemis and Brent, Mol. Cell Biol. 12:3006-3014, 1992).

LexA represents a preferred DNA binding domain in the invention. However, any other transcriptionally-inert or essentially transcriptionally-inert DNA binding domain may be used in the interaction trap system; such DNA binding domains are well known and include the DNA binding portions of the proteins ACE1 (CUP1), lambda cI, lac repressor, jun fos, or GCN4. For the above-described reasons, the GAL4 DNA binding domain represents a slightly less preferred DNA binding moiety for the bait proteins.

Bait proteins may be chosen from any protein of known or suspected diagnostic or therapeutic importance. Preferred bait proteins include oncoproteins (such as myc, particularly the C-terminus of myc, ras, src, fos, and particularly the oligomeric interaction domains of fos) or any other proteins involved in cell cycle regulation (such as kinases, phosphatases, the cytoplasmic portions of membrane-associated receptors, and other Cdc2 family members). In each case, the protein of diagnostic or therapeutic importance would be fused to a known DNA binding domain as generally described for LexA-Cdc2.

Reporters

As shown in Figure 3B, one preferred host strain according to the invention contains two different reporter genes, the *LEU2* gene and the *lacZ* gene, each carrying an upstream binding site for the bait protein. The reporter genes depicted in Figure 3B each include, as an upstream binding site, one or more LexA operators in place of their native Upstream Activation Sequences (UASs). These reporter genes may be integrated into the chromosome or may be carried on autonomously replicating plasmids (e.g., yeast 2μ plasmids).

A combination of two such reporters is pr ferred in the invention for a number of reasons. First, the *LexAop-LEU2* construction allows cells that contain interacting proteins to select themselves by growth on medium that lacks leucine, facilitating the examination of large numbers of potential interactor protein-containing cells. Second, the *LexAop-lacZ* reporter allows LEU^+ cells to be quickly screened to confirm an interaction. And, third, among other technical considerations (see below), the *LexAop-LEU2* reporter provides an extremely sensitive first selection, while the *LexAop-lacZ* reporter allows discrimination between proteins of different interaction affinities.

Although the reporter genes described herein represent a preferred embodiment of the invention, other equivalent genes whose expression may be detected or assayed by standard techniques may also be employed in conjunction with, or instead of, the *LEU2* and *lacZ* genes. Examples of other useful genes whose transcription can be detected include amino acid and nucleic acid biosynthetic genes (such as yeast *HIS3*, *URA3*, and *LYS2*) *GAL1*, *E. coli galK* (which complements the yeast *GAL1* gene), and the higher cell reporter genes CAT, GUS, and any gene encoding a cell surface antigen for which antibodies are available (e.g., CD4).

Prey proteins

In the selection described herein, a fourth DNA construction was utilized which encoded a series of candidate interacting proteins, each fused to a weak activation domain (i.e., prey proteins). One such prey protein construct is shown in Figure 3C; this plasmid encodes a prey fusion protein which includes an invariant N-terminal moiety. This moiety carries, amino to carboxy terminal, an ATG for protein expression, an optional nuclear localization s quence, a weak activation domain

(i.e., the B42 activation domain of Ma and Ptashne; Cell 51:113, 1987), and an optional epitope tag for rapid immunological detection of fusion protein synthesis. As described herein, a HeLa cDNA libraray was constructed, and random library sequences were inserted downstream of this N-terminal fragment to produce fusion genes encoding prey proteins.

Prey proteins other than those described herein are also useful in the invention. For example, cDNAs may be constructed from any mRNA population and inserted into an equivalent expression vector. Such a library of choice may be constructed *de novo* using commercially available kits (e.g., from Stratagene, La Jolla, CA) or using well established preparative procedures (see, e.g., *Current Protocols in Molecular Biology*, New York, John Wiley & Sons, 1987). Alternatively, a number of cDNA libraries (from a number of different organisms) are publically and commercially available; sources of libraries include, e.g., Clontech (Palo Alto, CA) and Stratagene (La Jolla, CA). It is also noted that prey proteins need not be naturally occurring full length polypeptides. For example, a prey protein may be encoded by a synthetic sequence or may be the product of a randomly generated open reading frame or a portion thereof. In one particular example, the prey protein includes only an interaction domain; such a domain may be useful as a therapeutic to modulate bait protein activity.

Similarly, other weak activation domains may be substituted for the B42 portion of the prey molecule; such activation domains must be weaker than the GAL4 activation region II moiety and preferably should be no stronger than B42 (as measured, e.g., by a comparison with GAL4 activation region II or B42 in parallel β-galactosidase assays using *lacZ* reporter genes); such a

domain may, however, be weaker than B42. In particular, the extraordinary sensitivity of the *LEU2* selection scheme (described above) allows even extremely weak activation domains to be utilized in the invention. Examples of other useful weak activation domains include B17, B112, and the amphipathic helix (AH) domains described in Ma and Ptashne (Cell 51:113, 1987), Ruden et al. (Nature 350:426-430, 1991), and Giniger and Ptashne (Nature 330:670, 1987).

Finally, the prey proteins, if desired, may include other optional nuclear localization sequences (e.g., those derived from the *GAL4* or *MATα2* genes) or other optional epitope tags (e.g., portions of the c-myc protein or the flag epitope available from Immunex). These sequences optimize the efficiency of the system, but are not absolutely required for its operation. In particular, the nuclear localization sequence optimizes the efficiency with which prey molecules reach the nuclear-localized reporter gene construct(s), thus increasing their effective concentration and allowing one to detect weaker protein interactions; and the epitope tag merely facilitates a simple immunoassay for fusion protein expression.

Those skilled in the art will also recognize that the above-described reporter gene, DNA binding domain, and gene activation domain components may be derived from any appropriate eukaryotic or prokaryotic source, including yeast, mammalian cell, and prokaryotic cell genomes or cDNAs as well as artificial sequences. Moreover, although yeast represents a preferred host organism for the interaction trap system (for reasons of ease of propagation, genetic manipulation, and large scale screening), other host organisms such as mammalian cells may also be utilized. If a mammalian system is chosen, a preferred reporter gene is the sensitive and

easily assayed CAT gene; useful DNA binding domains and gene activation domains may be chosen from those described above (e.g., the LexA DNA binding domain and the B42 or B112 activation domains).

The general type of interaction trap system described herein provides a number of advantages. For example, the system can be used to detect bait-prey interactions of varying affinity. This can be accomplished, e.g., by using reporter genes which differ quantitatively in their sensitivity to an interaction with a library protein. In particular, the equilibrium Kd with which a library-encoded protein must interact with the bait to activate the *LexAop-LEU2* reporter is probably $\leq 10^{-6}$M. This value is clearly sufficient to detect protein interactions that are weaker and shorter lived than those detected, e.g., by typical physical methods. The *lacZ* reporters are less sensitive, allowing the selection of different prey proteins by utilizing reporters with the appropriate number, affinity, and position of LexA operators; in particular, sensitivity of the *lacZ* reporter gene is increased by either increasing the number of upstream LexA operators, utilizing LexA operators which have increased affinity for LexA binding dimers, and/or decreasing the distance between the LexA operator and the downstream reporter gene promoter. This ability to manipulate the sensitivity of the system provides a measure of control over the strength of the interactions detected and thus increases the range of proteins which may be isolated.

The system provides at least three other advantages. First, the activation region on the library-encoded proteins is relatively weak, in order to avoid restrictions on the spectrum of library proteins detected; such restrictions are common when utilizing a strong, semi-toxic activation domain such as that of GAL4

or VP16 (Gill and Ptashne, Nature 334:721-724, 1988; Triezenberg et al., Genes Dev. 2:730-742, 1988; Berger et al., Cell 70:251-265, 1992). Second, the use of LexA to bind the bait to DNA allows the use of GAL4$^+$ yeast hosts and the use of the *GAL1* promoter to effect conditional expression of the library protein. This in turn allows the Leu or lacZ phenotypes to be unconditionally ascribed to expression of the library protein and minimizes the number of false positives; it also allows conditional expression and selection of interactor proteins which are toxic to the host cell if continuously produced. And third, placing the activation domain at the amino terminus, rather than at the carboxy terminus, of the fusion protein guarantees that the activation domain portion of the protein will be translated in frame, and therefore that one out of three fusion genes will encode a candidate activation domain-tagged interactor protein.

One particular interaction trap system is now described. The use of this system to isolate a protein (termed Cdi1) which physically interacts with a known cell division control protein (termed Cdc2) is also illustrated.

Isolation and Characterization of Cdi1

Isolation of the Cdi1 cDNA

To isolate proteins which interact with the cell division control protein Cdc2, the yeast strain EGY48/p1840 was utilized. This strain contained both the *LexAop-LEU2* and *LexAop-lacZ* reporters, as well as a plasmid that directed the synthesis of a LexA-Cdc2 bait protein (see below). The *LexAop-LEU2* reporter replaced the chromosomal *LEU2* gene. This reporter carried 3 copies of the high affinity colE1 double LexA operator (Ebina et al., J. Biol. Chem. 258:13258-13261, 1983) 40 nucleotides upstream of the major *LEU2* transcription startpoint. The *LexAop-lacZ* reporter (p1840) was carried

on a *URA3*$^+$ 2μ plasmid. This reporter carried a single LexA operator 167 nucleotides upstream of the major *GAL1* transcription startpoint.

A HeLa cDNA interaction library (described below) was also introduced into this strain using the plasmid depicted in Figure 3C (termed pJG4-5); this library vector was designed to direct the conditional expression of proteins under the control of a derivative of the *GAL1* promoter. This plasmid carried a 2μ replicator and a *TRP1*$^+$ selectable marker. cDNA was inserted into this plasmid on EcoR1-XhoI fragments. Downstream of the XhoI site, pJG4-5 contained the *ADH1* transcription terminator. The sequence of an invariant 107 amino acid moiety, encoded by the plasmid and fused to the N-terminus of all library proteins, is shown below the plasmid map in Figure 3C. This moiety carries, amino to carboxy terminal, an ATG, the SV40 T nuclear localization sequence (Kalderon et al., Cell 39:499-509, 1984), the B42 transcription activation domain, (Ma and Ptashne, Cell 51:113-119, 1987; Ruden et al., Nature 350:426-430, 1991) and the 12CA5 epitope tag from the influenza virus hemagglutinin protein (Green et al., Cell 28:477-487, 1982).

Following introduction of the prey-encoding plasmids into EGY48/p1840, over a million transformants were isolated, of which 3-4 X 10^5 expressed fusion proteins (see experimental procedures below). The colonies were pooled, diluted, and grown for five hours in liquid culture in the presence of galactose to induce synthesis of library-encoded proteins. The pool was then diluted again so that each original transformant was represented about 20 times and plated on galactose-containing medium without leucine. From about 2 X 10^7 cells, 412 LEU2$^+$ colonies were isolated. 55 of these colonies were blue on galactose Xgal medium, presumably

due to the lower sensitivity of the *lacZ* reporter. In all cells in which both reporters were active, both phenotypes were galactose-dependent, confirming that they required the library-encoded protein. Library plasmids were rescued from these cells, assigned to one of three classes by restriction mapping, and the plasmids identified from each class that contained the longest cDNA inserts. Synthesis of a fusion protein by the plasmid was verified in each case by Western blot analysis using anti-epitope antiserum.

Further analysis by detailed mapping and partial DNA sequencing showed that two of the recovered cDNA classes were identical to previously identified genes encoding CKS1hs and CKS2hs (Richardson et al., Genes Dev. 4:1332-1344, 1990), human homologs of the *S. pombe suc1*$^+$ product. Sequencing of the third restriction map class showed it to be a previously unidentified gene. This gene was termed *CDI1*, for Cdc2 Interactor 1; its protein product was termed Cdi1.

The *CDI1* gene was introduced into a panel of EGY48-derived strains (i.e., EGY48/1840 containing different LexA fusion baits) in order to test the reproducibility and specificity of the interaction between Cdc2 and Cdi1. Cells from 8 individual transformed cells that contained Cdi1 plus a given bait (horizontal streaks) or the same bait plus the library vector as a control (adjacent vertical streaks) were streaked with toothpicks onto each of three plates (Figure 4). The plates, shown in Figure 4, included a "control" plate, a Ura$^-$ Trp$^-$ His$^-$ glucose plate which selected for the presence of the bait plasmid, the *LexAop-lacZ* reporter, and the Cdi1 expression plasmid; a "glucose" plate, a Ura$^-$ Trp$^-$ His$^-$ Leu$^-$ glucose plate, which additionally selected for activation of the *LexAop-LEU2* reporter; and a "galactose" plate, a Ura$^-$ Trp$^-$ His$^-$

Leu⁻ galactose plate, which selected for the activation of the *LexAop-LEU2* reporter, and which induced the expression of Cdi1. Baits used in this test included: (1) LexA-Cdc2, (2) LexA-Bicoid, (3) LexA-Max, (4) LexA-Cln3, (5) LexA-Fus3, and (6) LexA-cMyc-Cterm (Figure 4).

As judged by the *LEU2* and *lacZ* transcription phenotypes, Cdi1 interacted specifically with LexA-Cdc2, and did not interact with LexA-cMyc-Cterm, LexA-Max, LexA-Bicoid, LexA-Cln3, or LexA-Fus3 (Figure 4). Cdi1 also interacted with other Cdc2 family proteins, including LexA-Cdc28, as discussed below. Applicants also note that, on glucose, the LexA-Cln3 bait weakly activated the *LexAop-LEU2* reporter, but that, on galactose, the inferiority of the carbon source and the dimished bait expression from the *ADH1* promoter eliminated this background.

The specificity of the Cdi1/Cdc2 interaction was then confirmed by physical criteria, in particular, by immunoprecipitation experiments. Extracts were made from EGY48 cells that contained a library plasmid that directed the synthesis of tagged Cdi1 and that also contained either a LexA-Cdc2 or a LexA-Bicoid bait.

In particular, 100 ml of cells were grown in glucose or galactose medium (in which Cdi1 expression was induced) to an OD_{600} of 0.6-0.8, pelleted by centrifugation, resuspended in 500µl RIPA, lysed by beating with glass beads five times for two minutes each, and spun twice for five minutes in a microfuge (10,000 X G) at 4° to remove the beads and cell debris. 5µl of this supernatant was taken as a control, and 15µl of rabbit anti-LexA antiserum was added to the remainder, which was incubated at 4°C for four hours on a rotating platform. LexA-containing proteins were first precipitated from this remainder with 50µl Staph A-coated

sepharose beads (Pharmacia, Piscataway, NJ) as described in Wittenberg and Reed (Cell 54:1061-1072, 1988). The entire pellet was then dissolved in Laemmli sample buffer, run on a 12.5% protein gel (SDS/PAGE), and blotted onto nitrocellulose. Tagged Cdi1 fusion proteins were identified by Western analysis of the blotted proteins with the 12CA5 monoclonal antihemagglutinin antibody essentially as described in Samson et al. (Cell 57:1045-1052, 1989).

The results are shown in Figure 5; the lanes are as follows: (1) Galactose medium, LexA-Bicoid bait, immunoprecipitation; (2) Glucose medium, LexA-Bicoid bait, immunoprecipitation; (3) Galactose medium, LexA-Bicoid bait, cell extract; (4) Glucose medium, LexA-Bicoid bait, cell extract; (5) Galactose medium, LexA-Cdc2 bait, immunoprecipitation; (6) Glucose medium, LexA-Cdc2 bait, immunoprecipitation; (7) Galactose medium, LexA-Cdc2 bait, cell extract; and (8) Glucose medium, LexA-Cdc2 bait, cell extract. As shown in Figure 5, anti-LexA antiserum precipitated Cdi1 from a yeast extract that contained LexA-Cdc2 and Cdi1, but not from one that contained LexA-Bicoid and Cdi1, thus confirming that Cdi1 physically interacted only with the Cdc2-containing bait protein.

<u>The Cdi1 Protein Product</u>

To analyze the Cdi1 protein product, the *Cdi1* cDNA was isolated from 12 different library plasmids that contained cDNAs of 4 different lengths. Sequence analysis revealed that all of the cDNA inserts contained an open reading frame, and inspection of the sequence of the longest cDNAs (Figure 6) revealed an ATG with a perfect match to the Kozak consensus translation initiation sequence (PuCC/GATGG) (Kozak, Cell 44:283-292, 1986). Careful analysis of the size of the Cdi1 mRNA in HeLa cells revealed that this ATG occurred between 15 and

45 nucleotides from the 5′ end of the Cdi1 message, suggesting that the longest cDNAs spanned the entire open reading frame.

The *Cdi1* gene is predicted to encode a protein of 212 amino acids. The Cdi1 amino acid sequence does not reveal compelling similarities to any previously identified proteins (Figure 6). However, two facts about the protein sequence are worth noting. First, 19 of the amino-terminal 35 amino acids are either proline, glutamic acid, serine, or threonine. Proteins that contain these stretches, called PEST sequences, are thought to be degraded rapidly (Rogers et al., Science 234:364-368, 1986); in fact, this stretch of Cdi1 is more enriched in these amino acids than the C-termini of the yeast G1 cyclins, in which the PEST sequences are known to be functional (Cross, Mol. Cell. Biol 8:4675-4684, 1988; Nash et al., EMBO J. 7:4335-4346, 1988; Hadwiger et al., Proc. Nat. Acad. Sci. USA 86:6255-6259, 1989). Second, since the cDNA library from which the plasmids that encoded Cdi1 were isolated was primed with oligo dT, and since all isolated Cdi1 cDNAs by definition encoded proteins that interacted with Cdc2, analysis of the sizes of Cdi1 cDNA inserts obtained in the screen necessarily localized the portion of the protein sufficient for interaction with Cdc2 to Cdi1's C-terminal ~170 amino acids.

Analysis of Cdi1 Function in Yeast

In initial efforts to understand Cdi1 function, the effects of Cdi1 expression in yeast were examined. In particular, because Cdi1 interacts with Cdc2 family proteins, including *S. cerevisiae* Cdc28, an examination of whether Cdi1 affected phenotypes that depended on other known proteins that interact with Cdc28 was undertaken.

Toward this end, the fact that expression of the *S. pombe suc1*$^{+}$ or *S. cerevisiae* Cks proteins can rescue the temperature sensitivity of strains that bear certain *cdc28*^{ts} alleles was exploited; this effect is thought to be due to the ability of these proteins to form complexes with the labile Cdc28^{ts} protein, protecting it against thermal denaturation (Hadwiger et al., Proc. Nat. Acad. Sci. USA 86:6255-6259, 1989). It was found that Cdi1 expression did not rescue the temperature-sensitivity of any *cdc28* allele tested, although human Cks2 did.

Next, the ability of Cdi1 to confer on yeast either of two phenotypes associated with expression of *S. cerevisiae* or higher eukaryotic cyclins was examined; such phenotypes include resistance to the arrest of *MATa* strains by α factor, and rescue of growth arrest of a strain deficient in Cln1, Cln2, and Cln3. Again, however, Cdi1 expression did not confer either phenotype.

During initial studies, it was noted that expression of Cdi1 inhibited yeast cell cycle progression. Cultures of cells that expressed Cdi1 increased their cell number and optical density more slowly than control populations (Figure 7A).

To further investigate this growth retardation phenotype, the morphology of Cdi1-expressing cells was examined. W303 cells were transformed with pJG4-4Cdi1, a galactose-inducible vector that directs the synthesis of Cdi1. Morphology of cells was examined with Nomarski optics at 1000X magnification. As shown in Figure 8, such microscopic examination of the cells showed that, compared with controls, cells in which Cdi1 was expressed were larger, and a subpopulation showed aberrant morphologies: 5% of the cells formed elongated schmoos, and 5% exhibited multiple buds. Immunofluorescent examination of a sample of these cells which had been

DAPI stained (as described below) showed that the nuclei of some of the largest cells were not condensed.

Finally, cells were examined for their ability to bud. Samples of 400 cells from control populations and from populations expressing Cdi1 were examined by phase contrast microscopy, and the budding index was calculated as the percentage of budded cells in each population as described in Wittenberg and Reed (Mol. Cell. Biol. 9:4064-4068, 1989). As shown in Figure 7B, less than 10% of the cells in the Cdi1-expressing population showed buds, as opposed to 30% of the cells in the control population, suggesting that fewer of the cells in the population expressing Cdi1 had passed through the G1 to S transition. This finding is consistent with the idea that the increased cell size and growth retardation were also due to a prolongation of G1.

This hypothesis was further tested by FACS analysis of cellular DNA. In particular, W303 cells that contained Cdi1 were grown as described above and diluted to OD_{600}=0.1 in 2% glucose or 1% raffinose, 1% galactose, and grown to OD_{600}=0.8-1.0. At this point, the cells were collected, sonicated, fixed in 70% ethanol, stained with propidium iodide, and subjected to FACS analysis to determine DNA content as previously described (Lew et al. Cell 63:317-328, 1992). Approximately 20,000 events were analyzed. These results, shown in Figure 7C, indicated that the majority of the cells in the Cdi1-expressing population had increased amounts of cellular DNA. This may indicate that an increased number of cells were in S phase; alternatively, it may simply be the result of larger cell size and increased quantity of mitochondrial DNA.

Taken together, these experiments thus indicated that protracted Cdi1 expression in *S. cerevisiae* caused a retardation in the passage of cells through the cell

cycle, most likely by increasing the proportion of cells in G1; they thus also indicate that Cdi1 expression uncoupled the normal synchrony between these two metrics of cell cycle progression.

Because Cdi1 interacts with Cdc2 family proteins, it was postulated that the Cdi1 growth retardation phenotype in *S. cerevisiae* might be explained by sequestration of Cdc28 into protein complexes that were not competent to cause the cell to traverse G1. To test this hypothesis, the effect of native Cdi1 expression in cells containing Cdc28 with and without overexpressed native human Cdc2 was compared. Specifically, W303 cells that carried the indicated combinations of galactose-inducible Cdi1 expression vector and/or Cdc2 expression vector were grown for 14h in complete minimal medium lacking tryptophan and histidine in the presence of 2% raffinose. Cells were then washed and diluted to OD_{600}=0.1 in the same media containing either 2% glucose, or 1% raffinose and 1% galactose. Optical densities were measured at two hour intervals for 12 hours. The results of these growth assay experiments are shown in Figure 7A.

Unexpectedly, it was found that the presence of additional Cdc2 increased the severity of the Cdi1-dependent growth inhibition (Figure 7A). This result suggested that Cdi1 endowed Cdc2 family proteins with a new function, at least in *S. cerevisiae*, one that inhibited their ability to cause cells to traverse G1 and S. The Cdi1 and Cdc2 expression plasmids together also caused some growth inhibition, even in glucose medium; this result was attributed to leaky expression from the *GAL1* promoter on the expression plasmid.

Analysis of Cdi1 Function in Mammalian Cells

The above results in yeast suggested that Cdi1 might have a similar effect on the ability of mammalian

cells to traverse G1 or S. Since Cdi1 was isolated from HeLa cDNA, the point in the cell cycle at which Cdi1 mRNA was expressed in these cells was first measured.

Specifically, adherent HeLa cells were synchronized in late G1 by a double thymidine block (Rao and Johnson, Nature 225:159-164, 1970) as described in Lew et al. (Cell 66:1197-1206, 1991). Aliquots of cells were collected every three hours after release from the block. Released cells reentered the cell cycle 9 hours after release, as measured by FACS analysis of DNA content. Total RNA was prepared from each aliquot at different time points, run out on a formaldehyde agarose gel, and blotted onto nylon (Nytran, Schleider and Schuell, Keene, NH) as described in Ausubel et al. (Current Protocols in Molecular Biology, New York, John Wiley & Sons, 1987). The blot was probed with random primed DNA probes (Feinberg and Vogelstein, Anal. Biochem. 132:6-13, 1983) made from a 690 bp EcoRI fragment that contained Cdi1, a 1389 bp PstI fragment from of human cyclin E sequence (Lew et al., Cell 66:1197-1206, 1991), a 1228bp NcoI-SphI fragment from the coding sequence of the human Cyclin B1 gene (Pines and Hunder, Cell 58:833-846, 1989), and a 1268bp PstI fragment carrying the full length human glyceraldehyde-phosphate-dehydrogenase (GAPD) gene (Tokunaga et al., Cancer Res. 47:5616-5619, 1987) which served as a normalization control. As is shown in Figure 9A, expression of Cdi1 mRNA peaks at the end of G1, immediately before the G1 to S transition, in parallel with the expression of the cyclin E message. This temporal expression pattern was consistent with the hypothesis that Cdi1 expression might affect the G1 to S transition.

To further test this idea, HeLa cells were transfected either with pBNCdi1, a construction that

directed the synthesis of Cdi1 under the control of the Moloney Murine Leukemia Virus LTR (see below), or with the vector alone. Individual transformed clones were selected by their resistance to G418, and cells from these clones were stained with propidium-iodide and subjected to FACS analysis to determine DNA content (as described below). The midpoint of G1 was defined as the mode of the distribution of each graph; the modes on the two panels were of different heights (272 counts for cells transformed with the vector, 101 counts for cells that contained Cdi1); this broadened peak in the Cdi1-expressing cells reflected the increased proportion of the population that contains approximately 1X DNA content. 4 independent transfectants were analysed; all yielded similar results. These results, which are shown in Figure 9B, indicated that the populations of cells in which Cdi1 was expressed contained an increased proportion of cells in G1 relative to control populations.

Cdc2-Cdi1 Interaction

To identify determinants of Cdc2 recognized by Cdi1, Cdi1 was tested for its ability to interact with a panel of different bait proteins that included Cdc2 proteins from yeast, humans, and flies, as well as the yeast Fus3 protein kinase (a protein kinase of the ERK class which negatively regulates Cln3 and which, by sequence criteria, is less related to the Cdc2 proteins than those proteins are to one another (Elion et al., Cell 60:649-664, 1990).

To perform these experiments, EGY48/JK103 (described below) containing a plasmid that directed the galactose-inducible synthesis of tagged Cdi1 was transformed with one of a series of different transcriptionally-inert LexA-Cdc2 family protein baits. Five individual transformants of each bait were grown to

OD_{600}=0.5-1.0 in minimal medium that contained 2% galactose but that lacked uracil, histidine, and tryptophan. Results are shown in Table 1 and are given in β-galactosidase units; variation among individual transformants was less than 20%.

TABLE 1

Bait	β-Galactosidase Activity
LexA-Cdc2 (Hs)	1580
LexA-Cdk2 (Hs)	440
LexA-Cdc28 (Sc)	480
LexA-Cdc2 (Dm)	40
LexA-Cdc2c (Dm)	>2
LexA-Fus3 (Sc)	>2

As shown in Table 1, tagged Cdi1 stimulated transcription from these baits to different levels; it activated strongly in strains that contained the human Cdc2 bait, against which it was selected, less strongly in strains that contained *S. cerevisiae* Cdc28 or human Cdk2 baits, and only weakly in strains that contained the DmCdc2 bait, one of the two *Drosophila* Cdc2 homologs (Jimenez et al., EMBO J. 9:3565-3571, 1990; Lehner and O'Farrell, EMBO J. 9:3573-3581, 1990). In strains that contained the DMCdc2c bait or Fus3, Cdi1 did not activate at all. Since baits in this panel were related in sequence, were made from the same vector, were translated from a message that had the same 5' untranslated sequence and the same LexA coding sequence, and were expressed in yeast in the same amounts, the differences in transcription among the bait strains very likely reflected differences in interaction with the tagged Cdi1.

In order to identify residues on Cdc2 proteins that Cdi1 might recognize, the transcription interaction data was compared to the sequence of the baits. A lineup

of the bait sequenc s was searched for residues that were conserved in the proteins with which Cdi1 interacted, but which differed in the proteins that Cdi1 did not touch. Use of this criterion identified 7 residues, which are indicated by asterisks in Figure 10. Of these residues, two, Glu 57 and Gly 154 (in human Cdc2), are altered in the non-interacting baits to amino acids of different chemical type. In DmCdc2c, residue 57 is changed from Glu to Asn, and residue 154 from Gly to Asn; in Fus3, these residues are changed to His and Asp. In human Cdc2, both of these residues adjoin regions of the molecule necessary for interaction with cyclins (Ducommun et al., Mol. Cell. Biol. 11:6177-6184, 1991). Projection of the human Cdc2 primary sequence on the crystal structure solved by Knighton et al. for bovine cAMP dependent protein kinase (Science 253:407-413, 1991) suggests that residues 57 and 154 are in fact likely to be close to these cyclin contact points in the folded protein.

These results are thus consistent with the idea that Cdi1 may exert its effects by changing the affinity of Cdc2 proteins for particular cyclins, thus potentially altering their substrate specificity.

In summary, Cdi1 is a protein which complexes with Cdc2 family proteins. It is expressed around the time of the G1 to S transition, and the above results suggest that it may negatively regulate passage of cells through this part of the cycle, thus linking the regulatory networks connecting extracellular signals with core cell cycle controls. If Cdi1 is in fact a negative regulator, it is interesting to note that its normal function may be to convey signals that retard or block the cell cycle during G1. Since both normal differentiation and cancer can be considered consequences of changes in G1 regulation, this idea raises the possibilities that Cdi1

may function to remove cells from active cycle to allow differentiation (Pardee, Science 246:603-608, 1989); and that there are cancers in which lesions in the G1 regulatory machinery prevent Cdi1 from exerting its full effect.

Experimental procedures

Bacteria and yeast

Manipulation of bacterial strains and of DNAs was by standard methods (see, e.g., Ausubel et al., Current Protocols in Molecular Biology, New York, John Wiley & Sons, 1987; and Sambrook et al., Molecular Cloning: a Laboratory Manual, Cold Spring Harbor, NY, Cold Spring Harbor Laboratory, 1989) unless otherwise noted. *E. coli* "Sure" *mcrA* Δ(*mrr*, *hsdRMS*, *mcrBC*) *endA1 supE44 thi-1 gyrA96 relA1 lac recB recJ sbcC umuC::Tn5(kan^R) uvrC* /F'[*proAB*, *lacI^{q}ZΔ_{M15}*]::Tn10(tet^R) (Stratagene Inc., LaJolla, CA) and KC8 (*pyrF::Tn5 hsdR leuB600 trpC9830 lacΔ74 strA galK hisB436*) were used as bacterial hosts throughout.

To determine whether Cdi1 complemented either G1 or G2 functions of *cdc28*, the following yeast strains were used: cdc28-1N (*MATa ura3 ade1 trp1 cdc28-1N*), which at the restrictive temperature arrests predominantly in G2; and cdc28-13 (*MATα leu2 trp1 his3 ura3 ade1 tyr1 cdc28-13*) and cdc28-17 (*MATa leu2 trp1 his3 ura3 met14 arg5 arg6 tyr1 cdc28-17*), which at the restrictive temperature arrests predominantly during G1.

Into these strains was introduced pJG4-6Cdi1 (see below), a yeast expression plasmid that directs the synthesis of Cdi1 that contains a hemagglutinin epitope tag at its amino terminus, and pJG4-7Cks2 (derived from the same selection) as a positive control. Overnight cultures of these strains were diluted 20:1 into trp$^-$ complete minimal medium with 2% glucose and 2% galactose

and grown at 25°C for five hours. Dilutions of these cultures were plated onto duplicate plates of solid media that contained the same carbon sources; one plate was placed at 25°C and the other at 36°C. Colonies were counted after five days of incubation

In order to determine whether Cdi1 complemented a strain deficient in G1 cyclins, strain 3c-1AX (*MATa bar1 Δcln1 Δcln2 Δcln3 cyh2 trp1 leu2 ura2 ade1 his2* [*pLEU2-CYH2* (*CYH*^{S})-*CLN3*$^+$]) into which pJG4-7Cdi1 or a GAL1-CLN3 construct as a positive control had been introduced was used. Overnight cultures were diluted into glucose and galactose medium as above, and grown for five hours at 30°C. Cells were plated onto glucose- and galactose-containing medium as above, except that the medium also contained 10μg/ml cyclohexamide; cells were grown for three days and counted. Colonies can only arise on this medium when the CYH^S-CLN3$^+$ plasmid is lost, an event which itself can only occur if the other plasmid rescues the Cln deficiency.

The ability of Cdi1 to cause resistance to arrest by α factor was tested using a derivative of W303 (*MATa trp1 ura3 his3 leu2 can1 bar1::LEU2*) into which pJG4-4Cdi1, a plasmid that directs the synthesis of native Cdi1, had been introduced. Strain W303 was also transformed with a set of mammalian cDNAs that had been isolated by their ability to confer α factor resistance as a positive control. Overnight cultures were grown in glucose and galactose as above, and then plated on glucose and galactose medium, in the presence and absence of 10^{-7}M α factor. Colonies were counted after 3 days.

For the growth rate experiments, W303 contained either pJG4-4Cdi1 or a vector control, in combination with either a pJG14-2, a *HIS3*$^+$ plasmid which directs the synthesis in yeast of native human Cdc2 under the control of the *ADH1* promoter, or a vector control. Overnight

cultures which were grown in His^- Trp^- minimal medium that contained 2% raffinose were collected, washed, and diluted into fresh medium that contained either 2% glucose or 1% galactose + 1% raffinose to OD_{600}=0.1. Growth kinetics were followed, measuring the OD of aliquots taken every 2 hours.

Baits

In order to optimize operator occupancy, baits were produced constitutively under the control of the *ADH1* promoter (Ammerer, Meth. Enzym. 101:192-210, 1983), and contained the LexA C-terminal oligomerization region, which contributes to operator occupancy by LexA-containing proteins, perhaps because it aids in the precise alignment of LexA amino termini of adjacent operator half sites (Golemis and Brent, Mol. Cell. Biol. 12:3006-3014, 1992). It is worth noting that all LexA-bait proteins so far examined enter the yeast nucleus in concentrations sufficient to permit operator binding, even though LexA derivatives are not specifically localized to the nucleus unless they contain other nuclear localization signals (see, e.g., Silver et al., Mol. Cell. Biol. 6:4763-4766, 1986).

pL202pl has been described (Ruden et al., Nature 350:426-430, 1991). This plasmid, a close relative of pMA424 and pSH2-1 (Ma and Ptashne, Cell 51:113-119, 1987; Hanes and Brent, Cell 57:1275-1283, 1989) carries the $HIS3^+$ marker and the 2μ replicator, and directs the synthesis in yeast of fusion proteins that carry the wild-type LexA protein at their amino terminus. Baits used in this study were made as follows: human Cdc2 (Lee and Nurse, Nature 327:31-35, 1987), Cdk2 (Tsai et al., Nature 353:174-177, 1991) and the *S. cerevisiae* CDC28 genes (Lorincz and Reed, Nature 307:183-185, 1984) were amplified by PCR using Vent polymerase (New England Biolabs, Beverley, MA) and cloned into pL202pl as EcoRI-

BamHI fragments. These proteins contained two amino acids (glu phe) inserted between the last amino acid of LexA and the bait proteins. The Drosophila Cdc2 (Jimenez et al., EMBO J. 9:3565-3571, 1990; Lehner and O'Farrell, EMBO J. 9:3573-3581, 1990) baits were cloned as BamHI-SalI fragments following PCR amplification. LexA-Fus3 (Elion, Cell 60:649-664, 1990) and LexA-Cln3 (Cross, Mol. Cell. Biol 8:4675-4684, 1988, Nash et al., EMBO J. 7:4335-4346, 1988) were made in a similar way except they were cloned as BamHI fragments. These plasmids contained five amino acids (glu phe pro gly ile) (SEQ ID NO:2) inserted between LexA and the baits. All these fusions contained the entire coding region from the second amino acid to the stop codon. LexA-cMyc-Cterm contained the carboxy-terminal 176 amino acids of human cMyc, and LexA-Max contained all of the human Max coding sequence. LexA-Bicoid (amino acid 2-160) has been described (Golemis and Brent, Mol. Cell. Biol. 12:3006-3014, 1992).

Reporters

In the interaction trap, one reporter, the *LexAop-LEU2* construction, replaced the yeast chromosomal *LEU2* gene. The other reporter, one of a series of *LexAop-GAL1-lacZ* genes (Brent and Ptashne, Cell 43:729-736, 1985; Kamens et al., Mol. Cell. Biol. 10:2840-2847, 1990), was carried on a 2μ plasmid. The reporters were designed so that their basal transcription was extremely low, presumably due both to the removal of the entirety of the UAS from both reporters, and to the fact (whose cause is unknown) that LexA operators introduced into promoters tend to decrease transcription (Brent and Ptashne, Nature 312:612-615, 1984; Lech, Gene activation by DNA-bound Fos and Myc proteins. Ph.D. thesis, Harvard University, 1990). Reporters were selected to differ in their response to activation by LexA fusion proteins. In this study, the *LEU2* reporter contained three copies of

the high-affinity LexA binding site found upstream of *E. coli* colE1 (Ebina et al., J. Biol. Chem. 258:13258-13261, 1983; Kamens et al., Mol. Cell. Biol. 10:2840-2847, 1990), and thus presumably binds a total of 6 dimers of the bait. In contrast, the *lacZ* gene employed in the primary screen contained a single lower affinity consensus operator (Brent and Ptashne, Nature 312:612-615, 1984) which binds a single dimer of the bait. The LexA operators in the *LEU2* reporter were closer to the transcription startpoint than they were in the *lacZ* reporter. These differences in the number, affinity, and position of the operators all contributed to making the *LEU2* gene a more sensitive indicator than the *lacZ* gene, a property that is useful for this method.

p1840 and pJK103 have been described (Brent and Ptashne, Cell 43:729-736, 1985, Kamens et al., Mol. Cell. Biol. 10:2840-2847, 1990). pHR33 (Ellerstrom et al., Plant Mol. Biol. 18:557-566, 1992) was cut with HindIII and an ~1166bp fragment that contained the *URA3*$^+$ gene from yEP24M13-2, a derivative of yEP24, was introduced into it to create pLEU2-0. This plasmid contains a BglII site 87 nucleotides upstream of the major *LEU2* transcription startpoint. pLEU2-0 was cut with BglII, and a 42bp double stranded BglII-ended oligomer

5'GATCCTGCTGTATATAAAACCAGTGGTTATATGTACAGTACG3'(SEQ ID NO 3)

3' GACGACATATATTTTGGTCACCAATATACATGTCATGCCTAG 5'(SEQ ID NO:4)

that contains the overlapping LexA operators found upstream of the colecin E1 gene (Ebina et al., J. Biol. Chem. 258:13258-13261, 1983) and which presumably binds 2 LexA dimers, was introduced into it. One plasmid, pLEU2-LexAop6, that contained three copies of this oligomer was picked; it presumably binds 6 dimers of LexA fusion proteins.

Selection strains

EGY12 (*MATa trp1 ura2 LEU2::pLEU2-0 (ΔUASLEU2)*) and EGY38 (as above but *::pLEU2-LexAop6*) were constructed as follows. pLEU2-0 and pLEU2-LexAop6 were linearized by digestion with ClaI within the *LEU2* gene, and the DNA was introduced into U457 (*MATa SUP53-a ade2-1 can1-100 ura3-52 trp1-1 [phi+]*) by lithium acetate transformation (Ito et al., J. Bacter. 153:163-168, 1983); ura^+ colonies, which presumably contained the plasmid DNA integrated into *LEU2*, were selected. Several of these transformants were grown in YPD. Ura^- cells were selected by plating these cultures on medium that contained 5-FOA (Ausubel et al., Current Protocols in Molecular Biology, New York, John Wiley & Sons, 1987). Both plasmids carry a TY1 element. For each integration, some of the $ura3^-$ revertants were also $trp1^-$, suggesting that the $URA3^+$ marker was deleted in a homologous recombination event that involved the TY1 sequences on the LEU2 plasmids and the chromosomal TY1 element upstream of *SUP53-a* (Oliver et al., Nature 357:38-46, 1992). Trp^- colonies from each integration, EGY12 (no LexA operators) and EGY38 (6 operators) were saved. These were mated to GG100-14D (*MATα his3 trp1 pho5*). The resulting diploids were sporulated, and a number of random (*MATα leu2- ura3- trp1- his3- GAL+*) spore products were recovered. EGY40 and EGY48 are products of this cross; EGY40 has no LexA operators, EGY48 has 6. To make the bait strains, EGY48 was transformed with p1840 or pJK103 and with the different bait plasmids. Double transformants were selected on Glucose Ura^- His^- plates, and expression of the bait protein confirmed by Western blotting using anti-LexA antibody and standard techniques.

Library ("prey") expression vectors

Library-encoded proteins were expressed from pJG4-5, a member of a series of expression plasmids designed

to be used in the interaction trap and to facilitate analysis of isolated proteins. These plasmids all carried the 2μ replicator, to ensure high copy number in yeast, and the *TRP1* marker. pJG4-5 was designed to possess the following features: a galactose-inducible promoter to allow conditional expression of the library proteins, an epitope tag to facilitate their detection, a nuclear localization signal to maximize their intranuclear concentration in order to increase the sensitivity of the selection, and a weak acid blob activation domain (Ma and Ptashne, Cell 51:113-119, 1987). This domain was chosen for two reasons: because its activity is not subject to known regulation by yeast proteins as is the major GAL4 activation domain, and, more importantly, because it is a weak activator, presumably avoiding toxicity due to squelching or other mechanisms (Gill and Ptashne, Nature 334:721-724, 1988, Berger et al., Cell 70:251-265, 1992) very likely to restrict the number or type of interacting proteins recovered.

pJG4-5 was constructed as follows. An "expression cassette" containing the GAL1 promoter and the ADH1 terminator and a 345 nt insert that encoded a 107 amino acid moiety was inserted into pJG4-0, a plasmid that carries the *TRP1* gene, the 2μ replicator, the pUC13 replication origin, and the ampicillin resistance gene. The pJG4-5 expression cassette directed the synthesis of fusion proteins, each of which carried at the amino terminus, amino to carboxy terminal, an ATG, an SV40 nuclear localization sequence (PPKKKRKVA) (SEQ ID NO: 5) (Kalderon et al., Cell 39:499-509, 1984), the B42 acid blob transcriptional activation domain (Ma and Ptashne, Cell 51:113-119, 1987) and the HA1 epitope tag (YPYDVPDYA) (SEQ ID NO: 6) (Green et al., Cell 28:477-487, 1980) (Figure 3C). In addition to this plasmid,

these experiments used two Cdi1 expression plasmids. EcoR1-XhoI Cdi1-containing fragments were introduced into pJG4-4 to make the plasmid pJG4-4Cdi1; Cdi1 was transcribed from this plasmid as a native, unfused protein under the control of the GAL1 promoter. EcoRI-XhoI Cdi1-containing fragments were also introduced into pJG4-6 to make the plasmid pJG4-6Cdi1; in this case, Cdi1 was expressed as an in-frame fusion containing, at its amino terminus, an ATG initiation codon and the hemagglutinin epitope tag.

Library construction

The activation-tagged yeast cDNA expression library was made from RNA isolated from serum grown, proliferating HeLa cells that were grown on plates to 70% confluence. Total RNA was extracted as described in Chomczynski and Sacchi (Anal. Biochem. 162:156-159, 1987), and polyA$^+$ mRNA was purified on an oligodT-cellulose column. cDNA synthesis was performed according to Gubler and Hoffman (Gene 25:263-269, 1983) as modified by Huse and Hansen (Strategies 1:1-3, 1988) using a linker primer that contained, 5′ to 3′, an 18nt polydT tract, an XhoI site, and a 25 nt long GA rich sequence to protect the XhoI site. To protect any internal XhoI sites, the first strand was synthesized in the presence of 5′-methyl-CTP (instead of CTP) with an RNAseH defective version of the Moloney virus reverse transcriptase (Superscript, BRL, Grand Island, NY). For second strand synthesis, the mRNA/cDNA hybrid was treated with RNAseH and *E. coli* DNA polymerase I, and the resulting ends were made flush by sequential treatment with Klenow, Mung Bean exonuclease, and Klenow onto which EcoRI adaptors:

5′ AATTCGGCACGAGGCG 3′ (SEQ ID NO: 7)
3′ GCCGTGCTCCGC 5′ (SEQ ID NO: 8)

were ligated, and the cDNA was digested with XhoI. This DNA was further purified on a Sephacryl S-400 spin column in order to remove excess adaptor sequences, and fractionated on a 5-20% KoAc gradient. Fractions containing >700 bp cDNAs were collected, and approximately 1/5 of the cDNA was ligated into EcoRI- and XhoI-digested pJG4-5. This ligation mixture was introduced into *E. coli* SURE cells by electroporation (Gene-Pulser, Bio-Rad, Hercules, CA) according to the manufacturer's instructions. 9.6×10^6 primary transformants were collected by scraping LB ampicillin plates. Colonies were pooled and grown in 6 liters of LB medium overnight (approximately three generations), and plasmid DNA was purified sequentially by standard techniques on two CsCl gradients. Digestion of transformants of individual library members with EcoR1 and XhoI revealed that >90% of the library members contained a cDNA insert whose typical size ranged between 1kb-2kb. Western blots of individual yeast transformants using the anti-hemagglutinin monoclonal antibody suggested that between 1/4 and 1/3 of the members expressed fusion proteins.

Selection of Cdc2 interactors

Library transformation of the above-described strain was performed according to the procedure described by Ito et al. (J. Bacter. 153:163-168, 1983), except that the cells were grown to a higher OD as described in Schiestl and Gietz (Curr. Genet 16:339-346, 1989) and single stranded carrier DNA was included in the transformation mix also as described in Schiestl and Gietz (Curr. Genet 16:339-346, 1989). This procedure gave 1.2×10^6 primary library transformants (10^4 library transformants/μg DNA). Transformants were selected on Glucose Ura^- His^- Trp^- plates, scraped, suspended in approximately 20 ml of 65% glycerol, 10mM Tris-HCl pH

7.5, 10mM $MgCl_2$, and stored in 1ml aliquots at -80°. Plating efficiency was determined on Galactose Ura$^-$ His$^-$ Trp$^-$ after growing 50µl of a cell suspension in 5 ml YP in the presence of 2% galactose. For screening the library, approximately 20 colony forming units on this medium/original transformant (about 2 X 10^7 cells) were plated on 4 standard circular 10cm Galactose Ura$^-$ His$^-$ Trp$^-$ Leu$^-$ plates after the YP/galactose induction described above.

412 Leu$^+$ colonies appeared after a 4 day incubation at 30°C. These colonies were collected on Glucose Ura$^-$ His$^-$ Trp$^-$ master plates and retested on Glucose Ura$^-$ His$^-$ Trp$^-$ Leu$^-$, Galactose Ura$^-$ His$^-$ Trp$^-$ Leu$^-$, Glucose Xgal Ura$^-$ His$^-$ Trp$^-$, and Galactose Xgal Ura$^-$ His$^-$ Trp$^-$ plates. 55 of these colonies showed galactose-dependent growth on leu$^-$ media and galactose-dependent blue color on Xgal medium, and were analyzed further.

Plasmid DNAs from these colonies were rescued as described (Hoffman and Winston, Gene 57:267-272, 1987), introduced into the bacterial strain KC8, and transformants were collected on Trp$^-$ ampicillin plates. Plasmid DNAs were analyzed and categorized by the pattern of restriction fragments they gave on 1.8% agarose 1/2X TBE gels after triple digestion with EcoRI and XhoI, and either AluI or HaeIII. Characteristic plasmids from different restriction map classes of these cDNAs were retransformed into derivatives of EGY48 that expressed a panel of different LexA fusion proteins. Plasmids that carried cDNAs whose encoded proteins interacted with the LexA-Cdc2 bait but not with other LexA fusion proteins, including LexA-Bicoid, LexA-Fus3, LexA-Cln3, LexA-cMyc-Cterm, and LexA-Max were characterized further.

Microscopy

5ml cultures of yeast cells were grown in the appropriate complete minimal medium up OD_{600} = 0.8-1 and

sonicated in a short burst to disrupt the clumps (Ausubel et al., Current Protocols in Molecular Biology, New York, John Wiley & Sons, 1987). The cells were collected by centrifugation, washed in 1ml TE, resuspended in 1ml 70% ethanol, and shaken for 1 hour at room temperature to fix them, then collected and resuspended in TE. The fixed cells were either examined directly at 1000x magnification with a Zeiss Axioscope microscope under Nomarski optics or by fluorescence after staining with 2.5μg/ml DAPI as described in Silver et al. (Mol. Cell. Biol. 6:4763-4766, 1986).

FACS analysis

Yeast cells were grown and fixed as described above and prepared for FACS analysis of DNA content essentially as in Lew et al. (Cell 63:317-328, 1992). After fixation the cells were collected and washed three times in 0.8 mls 50mM Tris/HCl pH 8.0, then 200μl 2mg/ml RNaseA was added and incubated at 37°C with continuous shaking for 5 hours. The cells were pelleted, resuspended in 0.5 ml of 5mg/ml pepsin (freshly dissolved in 55mM HCl) and incubated in a 37° waterbath for 30 minutes. The cells were spun down, washed with 1 ml of 200mM Tris/HCl pH 7.5, 211mM NaCl, 78mM $MgCl_2$ and resuspended in the same buffer. 55μl of 500 μg/ml propidium iodide was then added, and cells were stained overnight at 4°C. Typically 10,000-20,000 events were read and analysed in a Becton Dickinson Fluorescence Activated Cell Sorter (Becton Dickinson, Lincoln Park, NJ) with a CellFIT Cell-Cycle Analysis program Version 2.01.2.

For FACS analysis of DNA content, HeLa cells were grown on plates and transfected (Ausubel et al., Current Protocols in Molecular Biology, New York, John Wiley & Sons, 1987) either with pBNCdi1, a DNA copy of a retroviral cloning vector (Morgenstern and Land, Nucl.

Acids. Res. 18:3587-3596, 1990) that directs expression of native Cdi1 under the control of the MoMuLV promoter, or with the vector alone. Clones of transfected cells were selected by growth in medium that contained 400μg/ml of G418; Cdi1 expression did not diminish the number of G418 resistant cells recovered. Individual clones of each transfection (about 20) were rescued and grown on plates in DMEM + 10% calf serum, collected using 0.05% trypsin, 0.02% EDTA and washed once with 1X PBS. Cells from four clones derived from the Cdi1 transfection and four from the control transfection were suspended in 225μl of 30 μg/ml trypsin dissolved in 3.4mM citrate, 0.1% NP40, 1.5mM spermine and 0.5mM Tris, and incubated on a rotator for 10 minutes at room temperature. 188μl of 0.5mg/ml of trypsin inhibitor and 0.1 mg/ml RNAse A was then added and the suspension was vortexed. After adding 188μl of 0.4 mg/ml of propidium iodide and 1mg/ml spermine, the samples were incubated for 30 minutes at 4°C. FACS analysis was carried out as described above.

Cdi1 Polypeptides and Antibodies

Polypeptide Expression

In general, polypeptides according to the invention may be produced by transformation of a suitable host cell with all or part of a Cdi1-encoding cDNA fragment (e.g., the cDNA described above) in a suitable expression vehicle.

Those skilled in the field of molecular biology will understand that any of a wide variety of expression systems may be used to provide the recombinant protein. The precise host cell used is not critical to the invention. The Cdi1 polypeptide may be produced in a prokaryotic host (e.g., *E. coli*) or in a eukaryotic host (e.g., *Saccharomyces cerevisiae* or mammalian cells, e.g., COS 1, NIH 3T3, or HeLa cells). Such cells are available

from a wide range of sources (e.g., the American Type Culture Collection, Rockland, MD; also, see, e.g., Ausubel et al., *Current Protocols in Molecular Biology*, John Wiley & Sons, New York, 1989). The method of transformation or transfection and the choice of expression vehicle will depend on the host system selected. Transformation and transfection methods are described, e.g., in Ausubel et al. (*Current Protocols in Molecular Biology*, John Wiley & Sons, New York, 1989); expression vehicles may be chosen from those provided, e.g., in *Cloning Vectors*: A Laboratory Manual (P.H. Pouwels et al., 1985, Supp. 1987).

One preferred expression system is the mouse 3T3 fibroblast host cell transfected with a pMAMneo expression vector (Clontech, Palo Alto, CA). pMAMneo provides: an RSV-LTR enhancer linked to a dexamethasone-inducible MMTV-LTR promotor, an SV40 origin of replication which allows replication in mammalian systems, a selectable neomycin gene, and SV40 splicing and polyadenylation sites. DNA encoding a Cdi1 polypeptide would be inserted into the pMAMneo vector in an orientation designed to allow expression. The recombinant Cdi1 protein would be isolated as described below. Other preferable host cells which may be used in conjunction with the pMAMneo expression vehicle include COS cells and CHO cells (ATCC Accession Nos. CRL 1650 and CCL 61, respectively).

Alternatively, a Cdi1 polypeptide is produced by a stably-transfected mammalian cell line. A number of vectors suitable for stable transfection of mammalian cells are available to the public, e.g., see Pouwels et al. (supra); methods for constructing such cell lines are also publicly available, e.g., in Ausubel et al. (supra). In one example, cDNA encoding the Cdi1 polypeptide is cloned into an expression vector which includes the

dihydrofolate reductase (DHFR) gene. Integration of the plasmid and, therefore, the Cdi1-encoding gene into the host cell chromosome is selected for by inclusion of 0.01-300 μM methotrexate in the cell culture medium (as described in Ausubel et al., supra). This dominant selection can be accomplished in most cell types. Recombinant protein expression can be increased by DHFR-mediated amplification of the transfected gene. Methods for selecting cell lines bearing gene amplifications are described in Ausubel et al. (supra); such methods generally involve extended culture in medium containing gradually increasing levels of methotrexate. DHFR-containing expression vectors commonly used for this purpose include pCVSEII-DHRF and pAdD26SV(A) (described in Ausubel et al., supra). Any of the host cells described above or, preferably, a DHFR-deficient CHO cell line (e.g., CHO DHFR$^-$cells, ATCC Accession No. CRL 9096) are among the host cells preferred for DHFR selection of a stably-transfected cell line or DHFR-mediated gene amplification.

Once the recombinant Cdi1 protein is expressed, it is isolated, e.g., using affinity chromatography. In one example, an anti-Cdi1 antibody (e.g., produced as described herein) may be attached to a column and used to isolate the Cdi1 polypeptide. Lysis and fractionation of Cdi1-harboring cells prior to affinity chromatography may be performed by standard methods (see, e.g., Ausubel et al., supra). Alternatively, a Cdi1 fusion protein, for example, a Cdi1-maltose binding protein, a Cdi1-β-galactosidase, or a Cdi1-trpE fusion protein, may be constructed and used for isolation of Cdi1 protein (see, e.g., Ausubel et al., supra; New England Biolabs, Beverly, MA).

Once isolated, the recombinant protein can, if desired, be further purified, e.g., by high performance

liquid chromatography (see, e.g., Fisher, *Laboratory Techniques In Biochemistry And Molecular Biology*, eds., Work and Burdon, Elsevier, 1980).

Polypeptides of the invention, particularly short Cdi1 fragments, can also be produced by chemical synthesis (e.g., by the methods described in *Solid Phase Peptide Synthesis*, 2nd ed., 1984 The Pierce Chemical Co., Rockford, IL).

These general techniques of polypeptide expression and purification can also be used to produce and isolate useful Cdi1 fragments or analogs (described below).

Anti-Cdi1 Antibodies

Human Cdi1 (or immunogenic fragments or analogues) may be used to raise antibodies useful in the invention; such polypeptides may be produced by recombinant or peptide synthetic techniques (see, e.g., *Solid Phase Peptide Synthesis*, supra; Ausubel et al., supra). The peptides may be coupled to a carrier protein, such as KLH as described in Ausubel et al, supra. The KLH-peptide is mixed with Freund's adjuvant and injected into guinea pigs, rats, or preferably rabbits. Antibodies may be purified by peptide antigen affinity chromatography.

Monoclonal antibodies may be prepared using the Cdi1 polypeptides described above and standard hybridoma technology (see, e.g., Kohler et al., *Nature* 256:495, 1975; Kohler et al., *Eur. J. Immunol.* 6:511, 1976; Kohler et al., *Eur. J. Immunol.* 6:292, 1976; Hammerling et al., *In Monoclonal Antibodies and T Cell Hybridomas*, Elsevier, NY, 1981; Ausubel et al., supra).

Once produced, polyclonal or monoclonal antibodies are tested for specific Cdi1 recognition by Western blot or immunoprecipitation analysis (by the methods described in Ausubel et al., supra). Antibodies which specifically recognize a Cdi1 polypeptide are considered to be useful in the invention; such antibodies may be used, e.g., in

an immunoassay to monitor the level of Cdi1 produced by a mammal.

Therapeutic and Diagnostic Uses for the Cdi1 Polypeptide

Therapy

The Cdi1 polypeptide of the invention has been shown to interact with a key regulator of human cell division and to inhibit the *in vivo* proliferation of yeast and human cells. Because of its role in the control of cell division, Cdi1 is an unusually good candidate for an anti-cancer therapeutic. Preferably, this therapeutic is delivered as a sense or antisense RNA product, for example, by expression from a retroviral vector delivered, for example, to the bone marrow. Treatment may be combined with more traditional cancer therapies such as surgery, radiation, or other forms of chemotherapy.

Alternatively, using the interaction trap system described herein, a large number of potential drugs may be easily screened, e.g., in yeast, for those which increase or decrease the interaction between Cdi1 and Cdc2. Drugs which increase Cdc2:Cdi1 interaction would increase reporter gene expression in the instant system, and conversely drugs which decrease Cdc2:Cdi1 interaction would decrease reporter gene expression. Such drugs are then tested in animal models for efficacy and, if successful, may be used as anticancer therapeutics according to their normal dosage and route of administration.

Detection of A Malignant Condition

Cdi1 polypeptides may also find diagnostic use in the detection or monitoring of cancerous conditions. In particular, because Cdi1 is involved in the control of cell division, a change in the level of Cdi1 production may indicate a malignant or pre-malignant condition.

Levels of Cdi1 expression may be assayed by any standard technique. For example, its expression in a biological sample (e.g., a biopsy) may be monitored by standard Northern blot analysis or may be aided by PCR (see, e.g., Ausubel et al., supra; *PCR Technology: Principles and Applications for DNA Amplification*, ed., H.A. Ehrlich, Stockton Press, NY; and Yap and McGee, *Nucl. Acids. Res.* 19:4294, 1991). These techniques are enabled by the provision of the Cdi1 sequence.

Alternatively, immunoassays may be used to detect Cdi1 protein in a biological sample. Cdi1-specific polyclonal, or preferably monoclonal, antibodies (produced as described above) may be used in any standard immunoassay format (e.g., ELISA, Western blot, or RIA assay) to measure Cdi1 polypeptide levels; again comparison would be to wild type Cdi1 levels, and a change in Cdi1 production would be indicative of a malignant or pre-malignant condition. Examples of immunoassays are described, e.g., in Ausubel et al., supra. Immunohistochemical techniques may also be utilized for Cdi1 detection. For example, a tissue sample may be obtained from a patient, and a section stained for the presence of Cdi1 using an anti-Cdi1 antibody and any standard detection system (e.g., one which includes a secondary antibody conjugated to horseradish peroxidase). General guidance regarding such techniques can be found in, e.g., Bancroft and Stevens (*Theory and Practice of Histological Techniques*, Churchill Livingstone, 1982) and Ausubel et al. (supra).

In one particular example, a diagnostic method may be targeted toward a determination of whether the Cdi1 gene of a mammal includes the N-terminal PEST domain-encoding sequence. Because this sequence is very likely to stabilize the Cdi1 protein, its deletion may result in altered cellular levels of Cdi1 polypeptide and therefore

be indicative of a malignant or premalignant condition. PEST deletions may be identified either by standard nucleic acid or polypeptide analyses.

The Cdi1 polypeptide is also useful for identifying that compartment of a mammalian cell where important cell division control functions occur. Antibodies specific for Cdi1 may be produced as described above. The normal subcellular location of the protein is then determined either in situ or using fractionated cells by any standard immunological or immunohistochemical procedure (see, e.g., Ausubel et al., supra; Bancroft and Stevens, Theory and Practice of Histological Techniques, Churchill Livingstone, 1982).

The methods of the instant invention may be used to reduce or diagnose the disorders described herein in any mammal, for example, humans, domestic pets, or livestock. Where a non-human mammal is treated, the Cdi1 polypeptide or the antibody employed is preferably specific for that species.

Other Embodiments

In other embodiments, the invention includes any protein which is substantially homologous to human Cdi1 (Fig. 6, SEQ ID NO: 1); such homologs include other substantially pure naturally occurring mammalian Cdi1 proteins as well as allelic variations; natural mutants; induced mutants; proteins encoded by DNA that hybridizes to the Cdi1 sequence of Fig. 6 under high stringency conditions or low stringency conditions (e.g., washing at 2X SSC at 40°C with a probe length of at least 40 nucleotides); and polypeptides or proteins specifically bound by antisera directed to a Cdi1 polypeptide, especially by antisera to the active site or to the Cdc2 binding domain of Cdi1. The term also includes chimeric polypeptides that include a Cdi1 fragment.

The invention further includes analogs of any naturally occurring Cdi1 polypeptide. Analogs can differ from the naturally occurring Cdi1 protein by amino acid sequence differences, by post-translational modifications, or by both. Analogs of the invention will generally exhibit at least 70%, more preferably 80%, even more preferably 90%, and most preferably 95% or even 99%, homology with all or part of a naturally occurring Cdi1 sequence. The length of comparison sequences will be at least 8 amino acid residues, preferably at least 24 amino acid residues, and more preferably more than 35 amino acid residues. Modifications include *in vivo* and *in vitro* chemical derivatization of polypeptides, e.g., acetylation, carboxylation, phosphorylation, or glycosylation; such modifications may occur during polypeptide synthesis or processing or following treatment with isolated modifying enzymes. Analogs can also differ from the naturally occurring Cdi1 polypeptide by alterations in primary sequence. These include genetic variants, both natural and induced (for example, resulting from random mutagenesis by irradiation or exposure to ethanemethylsulfate or by site-specific mutagenesis as described in Sambrook, Fritsch and Maniatis, Molecular Cloning: A Laboratory Manual (2d ed.), CSH Press, 1989, hereby incorporated by reference; or Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, 1989, hereby incorporated by reference). Also included are cyclized peptides molecules and analogs which contain residues other than L-amino acids, e.g., D-amino acids or non-naturally occurring or synthetic amino acids, e.g., β or γ amino acids.

In addition to full-length polypeptides, the invention also includes Cdi1 polypeptide fragments. As used herein, the term "fragment", means at least 10

contiguous amino acids, preferably at least 30 contiguous amino acids, more preferably at least 50 contiguous amino acids, and most preferably at least 60 to 80 or more contiguous amino acids. Fragments of Cdi1 can be generated by methods known to those skilled in the art or may result from normal protein processing (e.g., removal of amino acids from the nascent polypeptide that are not required for biological activity or removal of amino acids by alternative mRNA splicing or alternative protein processing events).

Preferable fragments or analogs according to the invention are those which exhibit biological activity (for example, the ability to interfere with mammalian cell division as assayed herein). Preferably, a Cdi1 polypeptide, fragment, or analog exhibits at least 10%, more preferably 30%, and most preferably, 70% or more of the biological activity of a full length naturally occurring Cdi1 polypeptide.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Brent, Roger
Gyuris, Jeno
Golemis, Erica

(ii) TITLE OF INVENTION: Interaction Trap System for Isolating Novel Proteins

(iii) NUMBER OF SEQUENCES: 33

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Fish & Richardson
(B) STREET: 225 Franklin Street
(C) CITY: Boston
(D) STATE: Massachusetts
(E) COUNTRY: U.S.A.
(F) ZIP: 02110-2804

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: 3.5" Diskette, 1.44 Mb
(B) COMPUTER: IBM PS/2 Model 50Z or 55SX
(C) OPERATING SYSTEM: MS-DOS (Version 5.0)
(D) SOFTWARE: WordPerfect (Version 5.1)

(vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER: 07/969,038
(B) FILING DATE: 10/30/92
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER:
(B) FILING DATE:

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Clark, Paul T.
(B) REGISTRATION NUMBER: 30,162
(C) REFERENCE/DOCKET NUMBER: 00786/143001

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (617) 542-5070
(B) TELEFAX: (617) 542-8906
(C) TELEX: 200154

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 804
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

```
GGC ACT GGT CTC GAC GTG GGG CGG CCA GCG ATG GAG CCG CCC AGT TCA    48
Gly Thr Gly Leu Asp Val Gly Arg Pro Ala Met Glu Pro Pro Ser Ser
1               5                  10                  15

ATA CAA ACA AGT GAG TTT GAC TCA TCA GAT GAA GAG CCT ATT GAA GAT    96
Ile Gln Thr Ser Glu Phe Asp Ser Ser Asp Glu Glu Pro Ile Glu Asp
            20                  25                  30

GAA CAG ACT CCA ATT CAT ATA TCA TGG CTA TCT TTG TCA CGA GTG AAT   144
Glu Gln Thr Pro Ile His Ile Ser Trp Leu Ser Leu Ser Arg Val Asn
        35                  40                  45

TGT TCT CAG TTT CTC GGT TTA TGT GCT CTT CCA GGT TGT AAA TTT AAA   192
Cys Ser Gln Phe Leu Gly Leu Cys Ala Leu Pro Gly Cys Lys Phe Lys
    50                  55                  60

GAT GTT AGA AGA AAT GTC CAA AAA GAT ACA GAA GAA CTA AAG AGC TGT   240
Asp Val Arg Arg Asn Val Gln Lys Asp Thr Glu Glu Leu Lys Ser Cys
65                  70                  75                  80

GGT ATA CAA GAC ATA TTT GTT TTC TGC ACC AGA GGG GAA CTG TCA AAA   288
Gly Ile Gln Asp Ile Phe Val Phe Cys Thr Arg Gly Glu Leu Ser Lys
                85                  90                  95

TAT AGA GTC CCA AAC CTT CTG GAT CTC TAC CAG CAA TGT GGA ATT ATC   336
Tyr Arg Val Pro Asn Leu Leu Asp Leu Tyr Gln Gln Cys Gly Ile Ile
            100                 105                 110

ACC CAT CAT CAT CCA ATC GCA GAT GGA GGG ACT CCT GAC ATA GCC AGC   384
Thr His His His Pro Ile Ala Asp Gly Gly Thr Pro Asp Ile Ala Ser
        115                 120                 125

TGC TGT GAA ATA ATG GAA GAG CTT ACA ACC TGC CTT AAA AAT TAC CGA   432
Cys Cys Glu Ile Met Glu Glu Leu Thr Thr Cys Leu Lys Asn Tyr Arg
    130                 135                 140

AAA ACC TTA ATA CAC TGC TAT GGA GGA CTT GGG AGA TCT TGT CTT GTA   480
Lys Thr Leu Ile His Cys Tyr Gly Gly Leu Gly Arg Ser Cys Leu Val
145                 150                 155                 160

GCT GCT TGT CTC CTA CTA TAC CTG TCT GAC ACA ATA TCA CCA GAG CAA   528
Ala Ala Cys Leu Leu Leu Tyr Leu Ser Asp Thr Ile Ser Pro Glu Gln
                165                 170                 175

GCC ATA GAC AGC CTG CGA GAC CTA AGA GGA TCC GGG GCA ATA CAG ACC   576
Ala Ile Asp Ser Leu Arg Asp Leu Arg Gly Ser Gly Ala Ile Gln Thr
            180                 185                 190

ATC AAG CAA TAC AAT TAT CTT CAT GAG TTT CGG GAC AAA TTA GCT GCA   624
Ile Lys Gln Tyr Asn Tyr Leu His Glu Phe Arg Asp Lys Leu Ala Ala
        195                 200                 205

CAT CTA TCA TCA AGA GAT TCA CAA TCA AGA TCT GTA TCA AGA           666
His Leu Ser Ser Arg Asp Ser Gln Ser Arg Ser Val Ser Arg
    210                 215                 220

TAAAGGAATT CAAATAGCAT ATATATGACC ATGTCTGAAA TGTCAGTTCT CTAGCATAAT 726

TTGTATTGAA ATGAAACCAC CAGTGTTATC AACTTGAATG TAAATGTACA TGTGCAGATA 786
```

TTCCTAAAGT TTTATTGA 804

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 5
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

Glu Phe Pro Gly Ile
1 5

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 42
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GATCCTGCTG TATATAAAAC CAGTGGTTAT ATGTACAGTA CG 42

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 4:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 42
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

GACGACATAT ATTTTGGTCA CCAATATACA TGTCATGCCT AG 42

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 5:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 9
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

Pro Pro Lys Lys Lys Arg Lys Val Ala
1 5

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 6:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 9
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

Tyr Pro Tyr Asp Val Pro Asp Tyr Ala
1 5

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 7:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AATTCGGCAC GAGGCG 16

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 8:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 12
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

GCCGTGCTCC GC 12

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 9:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 73
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

Met Glu Asp Tyr Thr Lys Ile Glu Lys Ile Gly Glu Gly Thr Tyr Gly
1 5 10 15

Val Val Tyr Lys Gly Arg Lys Lys Thr Thr Gly Gln Val Val Ala Met
20 25 30

Lys Lys Ile Arg Leu Glu Ser Glu Glu Glu Gly Val Pro Ser Thr Ala
35 40 45

Ile Arg Glu Ile Ser Leu Leu Lys Glu Leu Arg His Pro Asn Ile Val
50 55 60

Ser Leu Gln Asp Val Leu Met Gln Asp

65 70

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 10:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 73
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

Met Glu Asn Phe Gln Lys Val Glu Lys Ile Gly Glu Gly Thr Tyr Gly
1 5 10 15

Val Val Tyr Lys Ala Arg Asn Lys Leu Thr Gly Glu Val Val Ala Leu
20 25 30

Lys Lys Ile Arg Leu Asp Thr Glu Thr Glu Gly Val Pro Ser Thr Ala
35 40 45

Ile Arg Glu Ile Ser Leu Leu Lys Glu Leu Asn His Pro Asn Ile Val
50 55 60

Lys Leu Leu Asp Val Ile His Thr Glu
65 70

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 11:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 82
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

Met Ser Gly Glu Leu Ala Asn Tyr Lys Arg Leu Glu Lys Val Gly Glu
1 5 10 15

Gly Thr Tyr Gly Val Val Tyr Lys Ala Leu Asp Leu Arg Pro Gly Gln
20 25 30

Gly Gln Arg Val Val Ala Leu Leu Lys Lys Ile Arg Leu Glu Ser Glu
35 40 45

Asp Glu Gly Val Pro Ser Thr Ala Ile Arg Glu Ile Ser Leu Leu Lys
50 55 60

Glu Leu Lys Asp Asp Asn Ile Val Arg Leu Tyr Asp Ile Val His Ser
65 70 75 80

Asp Ala

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 12:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 73
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

```
Met Glu Asp Phe Glu Lys Ile Glu Lys Ile Gly Glu Gly Thr Tyr Gly
 1               5                  10                  15

Val Val Tyr Lys Gly Arg Asn Arg Leu Thr Gly Gln Ile Val Ala Met
            20                  25                  30

Lys Lys Ile Arg Leu Glu Ser Asp Asp Glu Gly Val Pro Ser Thr Ala
        35                  40                  45

Ile Arg Glu Ile Ser Leu Leu Lys Glu Leu Lys His Glu Asn Ile Val
    50                  55                  60

Cys Leu Glu Asp Val Leu Met Glu Glu
65                  70
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 13:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 77
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

```
Met Thr Thr Ile Leu Asp Asn Phe Gln Arg Ala Glu Lys Ile Gly Glu
 1               5                  10                  15

Gly Thr Tyr Gly Ile Val Tyr Lys Ala Arg Ser Asn Ser Thr Gly Gln
            20                  25                  30

Asp Val Ala Leu Lys Lys Ile Arg Glu Leu Gly Glu Thr Glu Gly Val
        35                  40                  45

Pro Ser Thr Ala Ile Arg Glu Ile Ser Leu Leu Lys Asn Leu Lys His
    50                  55                  60

Pro Asn Val Val Gln Leu Phe Asp Val Val Ile Ser Gly
65                  70                  75
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 86
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

Met Pro Lys Arg Ile Val Tyr Asn Ile Ser Ser Asp Phe Gln Leu Lys
1 5 10 15

Ser Leu Leu Gly Glu Gly Ala Tyr Gly Val Val Cys Ser Ala Thr His
20 25 30

Lys Pro Thr Gly Glu Ile Val Ala Ile Lys Lys Ile Glu Pro Phe Asp
35 40 45

Lys Pro Leu Phe Ala Leu Arg Thr Leu Arg Glu Ile Lys Ile Leu Lys
50 55 60

His Phe Lys His Glu Asn Ile Ile Thr Ile Phe Asn Ile Gln Arg Pro
65 70 75 80

Asp Ser Phe Glu Asn Phe
85

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 15:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 84
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

Ser Arg Leu Tyr Leu Ile Phe Glu Phe Leu Ser Met Asp Leu Lys Lys
1 5 10 15

Tyr Leu Asp Ser Ile Pro Pro Gly Gln Tyr Met Asp Ser Ser Leu Val
20 25 30

Lys Ser Tyr Leu Tyr Gln Ile Leu Gln Gly Ile Val Phe Cys His Ser
35 40 45

Arg Arg Val Leu His Arg Asp Leu Lys Pro Gln Asn Leu Leu Ile Asp
50 55 60

Asp Lys Gly Thr Ile Lys Leu Ala Asp Phe Gly Leu Ala Arg Ala Phe
65 70 75 80

Gly Ile Pro Ile

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 16:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

```
Asn Lys Leu Tyr Leu Val Phe Glu Phe Leu His Gln Asp Leu Lys Lys
 1               5                  10                  15

Phe Met Asp Ala Ser Ala Leu Thr Gly Ile Pro Leu Pro Leu Ile Lys
            20                  25                  30

Ser Tyr Leu Phe Gln Leu Leu Gln Gly Leu Ala Pro Cys His Ser His
        35                  40                  45

Arg Val Leu His Arg Asp Leu Lys Pro Gln Asn Leu Leu Ile Asn Thr
    50                  55                  60

Glu Gly Ala Ile Lys Leu Ala Asp Phe Gly Leu Ala Arg Ala Phe Gly
65                  70                  75                  80

Val Pro Val
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 17:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 84
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

```
His Lys Leu Tyr Leu Val Phe Glu Phe Leu Asp Leu Asp Leu Lys Arg
 1               5                  10                  15

Tyr Met Glu Gly Ile Pro Lys Asp Gln Pro Leu Gly Ala Asp Ile Val
            20                  25                  30

Lys Lys Phe Met Met Gln Leu Cys Lys Gly Ile Ala Tyr Cys His Ser
        35                  40                  45

His Arg Ile Leu His Arg Asp Leu Lys Pro Gln Asn Leu Leu Ile Asn
    50                  55                  60

Lys Asp Gly Asn Leu Lys Leu Gly Asp Phe Gly Leu Ala Arg Ala Phe
65                  70                  75                  80

Gly Val Pro Leu
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 18:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 84
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18:

```
Asn Arg Ile Tyr Leu Ile Phe Glu Phe Leu Ser Met Asp Leu Lys Lys
 1               5                  10                  15

Tyr Met Asp Ser Leu Pro Val Asp Lys His Met Glu Ser Glu Leu Val
            20                  25                  30

Arg Ser Tyr Leu Tyr Gln Ile Thr Ser Ala Ile Leu Phe Cys His Arg
        35                  40                  45

Arg Arg Val Leu His Arg Asp Leu Lys Pro Gln Asn Leu Leu Ile Asp
    50                  55                  60

Lys Ser Gly Leu Ile Lys Val Ala Asp Phe Gly Leu Gly Arg Ser Phe
65                  70                  75                  80

Gly Ile Pro Val
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 19:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 82
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 19:

```
Asn Asn Leu Tyr Met Ile Phe Glu Tyr Leu Asn Met Asp Leu Lys Lys
 1               5                  10                  15

Leu Met Asp Lys Lys Lys Asp Val Phe Thr Pro Gln Leu Ile Lys Ser
            20                  25                  30

Tyr Met His Gln Ile Leu Asp Ala Val Gly Phe Cys His Thr Asn Arg
        35                  40                  45

Ile Leu His Arg Asp Leu Lys Pro Gln Asn Leu Leu Val Asp Thr Ala
    50                  55                  60

Gly Lys Ile Lys Leu Ala Asp Phe Gly Leu Ala Arg Ile Phe Asn Val
65                  70                  75                  80

Pro Met
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 20:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 86
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

```
Asn Glu Val Tyr Ile Ile Gln Glu Leu Met Gln Thr Asp Leu His Arg
 1               5                  10                  15

Val Ile Ser Thr Gln Met Leu Ser Asp Asp His Ile Gln Tyr Phe Ile
            20                  25                  30

Tyr Gln Thr Leu Arg Ala Val Lys Val Leu Glu Gly Ser Asn Val Ile
        35                  40                  45

His Arg Asp Leu Lys Pro Ser Asn Leu Leu Ile Asn Ser Asn Cys Asp
    50                  55                  60

Leu Lys Val Cys Asp Phe Gly Leu Ala Arg Ile Ile Asp Glu Ser Ala
65                  70                  75                  80

Ala Asp Asn Ser Glu Pro
                85
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 21:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:

```
Arg Val Tyr Thr His Glu Val Val Thr Leu Trp Tyr Arg Ser Pro Glu
 1               5                  10                  15

Val Leu Leu Gly Ser Ala Arg Tyr Ser Thr Pro Val Asp Ile Trp Ser
            20                  25                  30

Ile Gly Thr Ile Phe Ala Glu Leu Ala Thr Lys Lys Pro Leu Phe His
        35                  40                  45

Gly Asp Ser Glu Ile Asp Gln Leu Phe Arg Ile Phe Arg Ala Leu Gly
    50                  55                  60

Thr Pro Asn Asn Glu Val Trp Pro Glu Val Glu Ser Leu Gln Asp Tyr
65                  70                  75                  80

Lys Asn Thr
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 22:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:

```
Arg Thr Tyr Thr His Glu Val Val Thr Leu Trp Tyr Arg Ala Pro Glu
 1               5                  10                  15

Ile Leu Leu Gly Cys Lys Tyr Tyr Ser Thr Ala Val Asp Ile Trp Ser
            20                  25                  30

Leu Gly Cys Ile Phe Ala Glu Met Val Thr Arg Arg Ala Leu Phe Pro
        35                  40                  45

Gly Asp Ser Glu Ile Asp Gln Leu Phe Arg Ile Phe Arg Thr Leu Gly
    50                  55                  60

Thr Pro Asp Glu Val Val Trp Pro Gly Val Thr Ser Met Pro Asp Tyr
65                  70                  75                  80

Lys Pro Ser
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 23:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:

```
Arg Ala Tyr Thr His Glu Ile Val Thr Leu Trp Tyr Arg Ala Pro Glu
 1               5                  10                  15

Val Leu Leu Gly Gly Lys Gln Tyr Ser Thr Gly Val Asp Thr Trp Ser
            20                  25                  30

Ile Gly Cys Ile Phe Ala Glu Met Cys Asn Arg Lys Pro Ile Phe Ser
        35                  40                  45

Gly Asp Ser Glu Ile Asp Gln Leu Phe Lys Ile Phe Arg Val Leu Gly
    50                  55                  60

Thr Pro Asn Glu Ala Ile Trp Pro Asp Ile Val Tyr Leu Pro Asp Phe
65                  70                  75                  80

Lys Pro Ser
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 24:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:

```
Arg Ile Tyr Thr His Glu Ile Val Thr Leu Trp Tyr Arg Ala Pro Glu
 1               5                  10                  15

Val Leu Leu Gly Ser Pro Arg Tyr Ser Cys Pro Val Asp Ile Trp Ser
            20                  25                  30

Ile Gly Cys Ile Phe Ala Glu Met Ala Thr Arg Lys Pro Leu Phe Gln
        35                  40                  45

Gly Asp Ser Glu Ile Asp Gln Leu Phe Lys Ile Phe Arg Val Leu Gly
    50                  55                  60

Thr Pro Asn Glu Ala Ile Trp Pro Asp Ile Val Tyr Leu Pro Asp Phe
65                  70                  75                  80
Lys Pro Ser
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 25:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 83
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

```
Arg Ala Tyr Thr His Glu Val Val Thr Leu Trp Tyr Arg Ala Pro Glu
 1               5                  10                  15

Ile Leu Leu Gly Thr Lys Phe Tyr Ser Thr Gly Val Asp Ile Trp Ser
            20                  25                  30

Leu Gly Cys Ile Phe Ser Glu Met Ile Met Arg Arg Ser Leu Phe Pro
        35                  40                  45

Gly Asp Ser Glu Ile Asp Gln Leu Tyr Arg Ile Phe Arg Thr Leu Ser
    50                  55                  60

Thr Pro Asp Glu Thr Asn Trp Pro Gly Val Thr Gln Leu Pro Asp Phe
65                  70                  75                  80

Lys Thr Lys
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 26:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 90
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 26:

```
Thr Gly Gln Gln Ser Gly Met Thr Glu Tyr Val Ala Thr Arg Trp Tyr
 1               5                  10                  15

Arg Ala Pro Glu Val Met Leu Thr Ser Ala Lys Tyr Ser Arg Ala Met
            20                  25                  30

Asp Val Trp Ser Cys Gly Cys Ile Leu Ala Glu Leu Phe Leu Arg Arg
        35                  40                  45

Pro Ile Phe Pro Gly Arg Asp Tyr Arg His Gln Leu Leu Leu Ile Phe
    50                  55                  60

Gly Ile Ile Gly Thr Pro His Ser Asp Asn Asp Leu Arg Cys Ile Glu
65                  70                  75                  80

Ser Pro Arg Ala Arg Glu Tyr Ile Lys Ser
                85                  90
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 27:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

```
Phe Pro Lys Trp Lys Pro Gly Ser Leu Ala Ser His Val Lys Asn Leu
 1               5                  10                  15

Asp Glu Asn Gly Leu Asp Leu Leu Ser Lys Met Leu Ile Tyr Asp Pro
            20                  25                  30

Ala Lys Arg Ile Ser Gly Lys Met Ala Leu Asn His Pro Tyr Phe Asn
        35                  40                  45

Asp Leu Asp Asn Gln Ile Lys Lys Met
    50                  55
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 28:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 59
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 28:

```
Phe Pro Lys Trp Ala Arg Gln Asp Phe Ser Lys Val Val Pro Pro Leu
 1               5                  10                  15

Asp Glu Asp Gly Ile Asp Leu Leu Asp Lys Leu Leu Ala Tyr Asp Pro
            20                  25                  30
```

Asn Lys Arg Ile Ser Ala Lys Ala Ala Leu Ala His Pro Phe Thr Gln
35 40 45

Asp Val Thr Lys Pro Val Pro His Leu Arg Leu
50 55

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 29:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 29:

Phe Pro Gln Trp Arg Arg Lys Asp Leu Ser Asn Gln Leu Lys Asn Leu
1 5 10 15

Asp Ala Asn Gly Ile Asp Leu Ile Gln Lys Met Leu Ile Tyr Asp Pro
20 25 30

Val His Arg Ile Ser Ala Lys Asp Ile Leu Glu His Pro Tyr Phe Asn
35 40 45

Gly Phe Gln Ser Gly Leu Val Arg Asn
50 55

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 30:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 57
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 30:

Phe Pro Gln Trp Arg Arg Lys Asp Leu Ser Asn Gln Leu Lys Asn Leu
1 5 10 15

Asp Ala Asn Gly Ile Asp Leu Ile Gln Lys Met Leu Ile Tyr Asp Pro
20 25 30

Val His Arg Ile Ser Ala Lys Asp Ile Leu Glu His Pro Tyr Phe Asn
35 40 45

Gly Phe Gln Ser Gly Leu Val Arg Asn
50 55

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 31:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 72
(B) TYPE: amino acid
(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 31:

```
Phe Pro Arg Trp Glu Gly Thr Asn Met Pro Gln Pro Ile Thr Glu His
 1               5                  10                  15

Glu Ala His Glu Leu Ile Met Ser Met Leu Cys Tyr Asp Pro Asn Leu
            20                  25                  30

Arg Ile Ser Ala Lys Asp Ala Leu Gln His Ala Tyr Phe Arg Asn Val
        35                  40                  45

Gln His Val Asp His Val Ala Leu Pro Val Asp Pro Asn Ala Gly Ser
    50                  55                  60

Ala Ser Arg Leu Thr Arg Leu Val
65                  70
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 32:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 60
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 32:

```
Leu Pro Met Tyr Pro Ala Ala Pro Leu Glu Lys Met Phe Pro Arg Val
 1               5                  10                  15

Asn Pro Lys Gly Ile Asp Leu Leu Gln Arg Met Leu Val Phe Asp Pro
            20                  25                  30

Ala Lys Arg Ile Thr Ala Lys Glu Ala Leu Glu His Pro Tyr Leu Gln
        35                  40                  45

Thr Tyr His Asp Pro Asn Asp Glu Pro Glu Gly Glu
    50                  55                  60
```

(2) INFORMATION FOR SEQUENCE IDENTIFICATION NUMBER: 33:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 345
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 33:

```
AAG CTT ATG GGT GCT CCT CCA AAA AAG AAG AGA AAG GTA GCT GGT ATC    48
Lys Leu Met Gly Ala Pro Pro Lys Lys Lys Arg Lys Val Ala Gly Ile
1               5                  10                  15
```

```
AAT AAA GAT ATC GAG GAG TGC AAT GCC ATC ATT GAG CAG TTT ATC GAC     96
Asn Lys Asp Ile Glu Glu Cys Asn Ala Ile Ile Glu Gln Phe Ile Asp
            20                  25                  30

TAC CTG CGC ACC GGA CAG GAG ATG CCG ATG GAA ATG GCG GAT CAG GCG    144
Tyr Leu Arg Thr Gly Gln Glu Met Pro Met Glu Met Ala Asp Gln Ala
        35                  40                  45

ATT AAC GTG GTG CCG GGC ATG ACG CCG AAA ACC ATT CTT CAC GCC GGG    192
Ile Asn Val Val Pro Gly Met Thr Pro Lys Thr Ile Leu His Ala Gly
    50                  55                  60

CCG CCG ATC CAG CCT GAC TGG CTG AAA TCG AAT GGT TTT CAT GAA ATT    240
Pro Pro Ile Gln Pro Asp Trp Leu Lys Ser Asn Gly Phe His Glu Ile
65                  70                  75                  80

GAA GCG GAT GTT AAC GAT ACC AGC CTC TTG CTG AGT GGA GAT GCC TCC    288
Glu Ala Asp Val Asn Asp Thr Ser Leu Leu Leu Ser Gly Asp Ala Ser
                85                  90                  95

TAC CCT TAT GAT GTG CCA GAT TAT GCC TCT CCC GAA TTC GGC CGA CTC    336
Tyr Pro Tyr Asp Val Pro Asp Tyr Ala Ser Pro Glu Phe Gly Arg Leu
            100                 105                 110

GAG AAG CTT                                                       345
Glu Lys Leu
        115
```

Claims

1. A method for determining whether a first protein is capable of physically interacting with a second protein, comprising:

(a) providing a host cell which contains

(i) a reporter gene operably linked to a protein binding site;

(ii) a first fusion gene which expresses a first fusion protein, said first fusion protein comprising said first protein covalently bonded to a binding moiety which is capable of specifically binding to said protein binding site; and

(iii) a second fusion gene which expresses a second fusion protein, said second fusion protein comprising said second protein covalently bonded to a weak gene activating moiety; and

(b) measuring expression of said reporter gene as a measure of an interaction between said first and said second proteins.

2. The method of claim 1, further comprising isolating the gene encoding said second protein.

3. The method of claim 1, wherein said weak gene activating moiety is of lesser activation potential than GAL4 activation region II.

4. The method of claim 3, wherein said weak gene activating moiety is the B42 activation domain.

5. The method of claim 1, wherein said host cell is a yeast cell.

6. The method of claim 1, wherein said reporter gene comprises the *LEU2* gene or the *lacZ* gene.

7. The method of claim 1, wherein said host cell further contains a second reporter gene operably linked to said protein binding site.

8. The method of claim 1, wherein said protein binding site is a LexA binding site and said binding moiety comprises a LexA DNA binding domain.

9. The method of claim 1, wherein said second protein is a protein involved in the control of eukaryotic cell division.

10. The method of claim 9, wherein said cell division control protein is encoded by a Cdc2 gene.

11. A substantially pure preparation of Cdi1 polypeptide.

12. The polypeptide of claim 11, comprising an amino acid sequence substantially identical to the amino acid sequence shown in Figure 6 (SEQ ID NO: 1).

13. Purified DNA comprising a sequence encoding a polypeptide of claims 11 or 12.

14. The purified DNA of claim 13, wherein said DNA is cDNA.

15. The purified DNA of claim 11, wherein said DNA encodes a human Cdi1 polypeptide.

16. A vector comprising the purified DNA of claim 15.

17. A cell containing the purified DNA of claim 15.

18. A method of producing a recombinant Cdi1 polypeptide comprising,

providing a cell transformed with DNA encoding a Cdi1 polypeptide positioned for expression in said cell;

culturing said transformed cell under conditions for expressing said DNA; and

isolating said recombinant Cdi1 polypeptide.

19. A purified antibody which binds specifically to a polypeptide of claims 11 or 12.

20. A method of detecting a malignant cell in a biological sample, said method comprising measuring Cdi1 gene expression in said sample, a change in Cdi1 expression relative to a wild-type sample being indicative of the presence of said malignant cell.

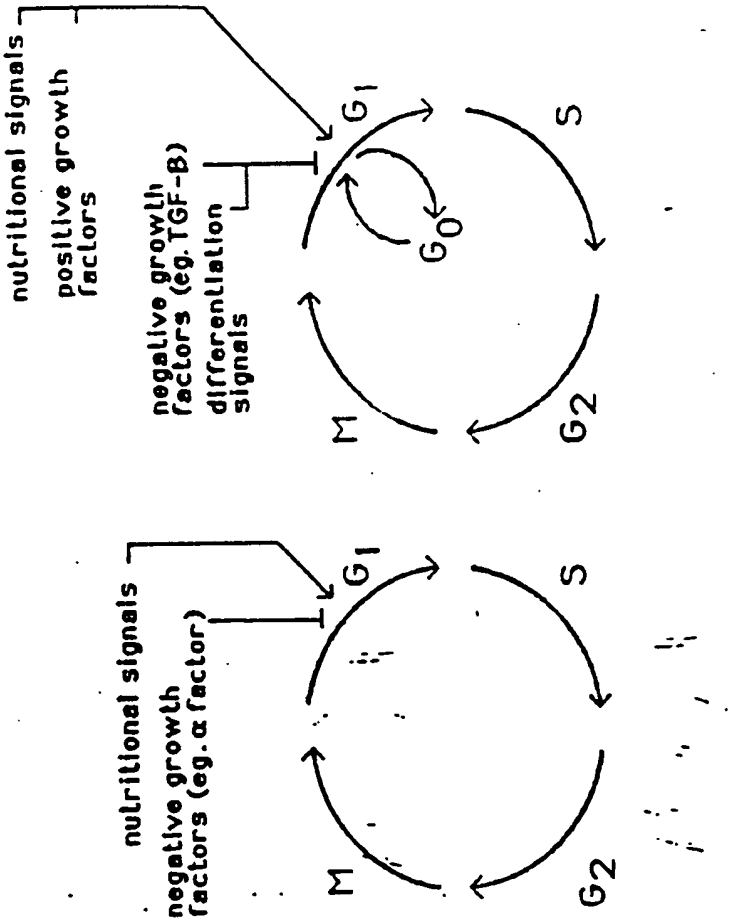

nutritional signals
positive growth factors
negative growth factors (eg. TGF-β)
differentiation signals
G1
S
G0
M
G2
nutritional signals
negative growth factors (eg. α factor)
G1
S
M
G2

Fig. 1B

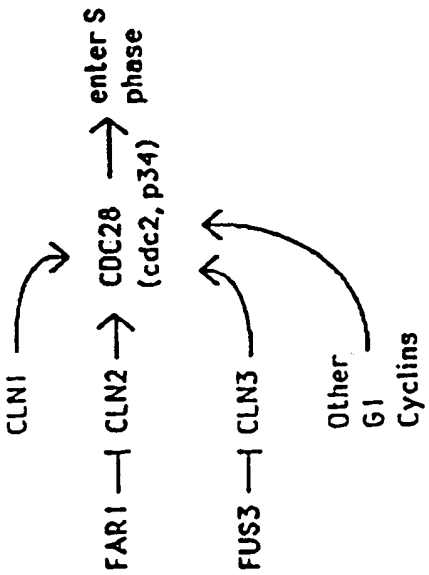

CLN1
CLN2
FAR1
CDC28
(cdc2, p34)
enter S phase
FUS3
CLN3
Other G1 Cyclins

Fig. 1A

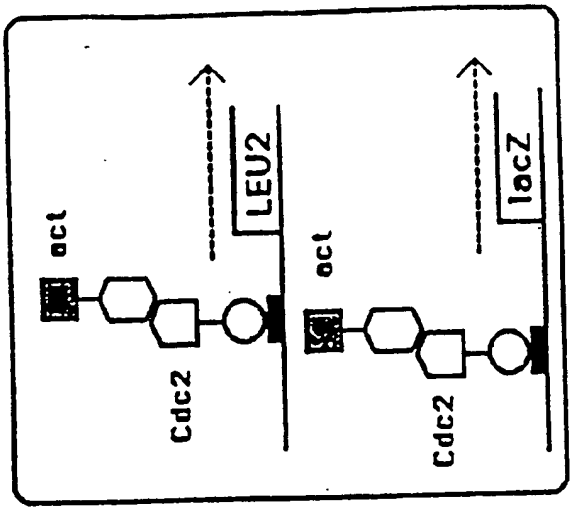


Fig. 2C

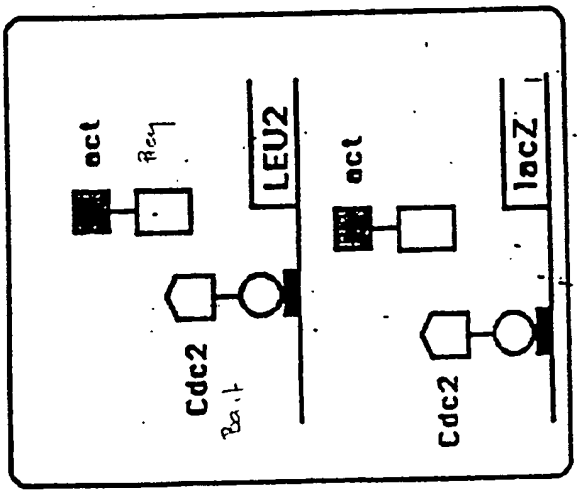


Fig. 2B

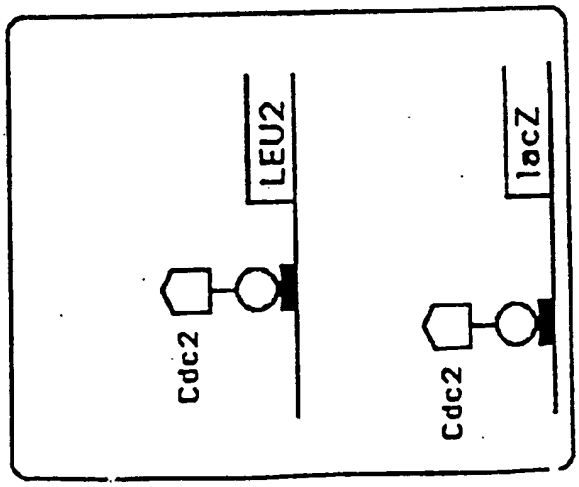
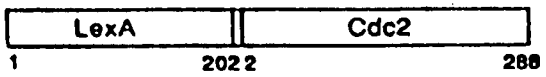


Fig. 2A



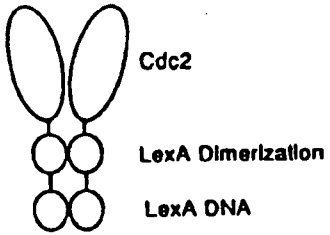


Fig. 3A

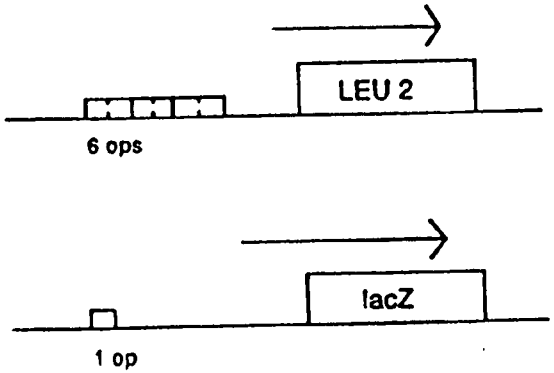
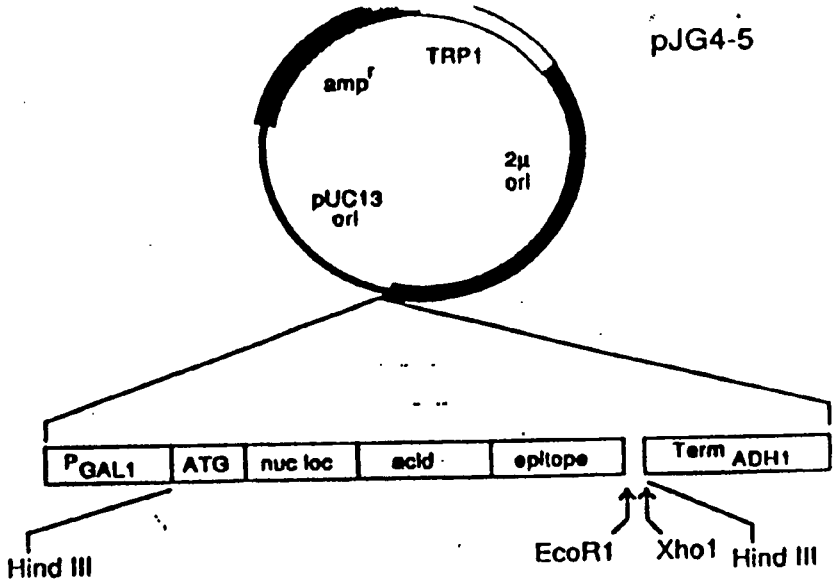


Fig. 3B




```
1    AAG CTT ATG GGT GCT CCT CCA AAA AAG AAG AGA AAG GTA GCT GGT
1          M   G   A   P   P   K   K   K   R   K   V   A   G

46   ATC AAT AAA GAT ATC GAG GAG TGC AAT GCC ATC ATT GAG CAG TTT
14   I   N   K   D   I   E   E   C   N   A   I   I   E   Q   F

91   ATC GAC TAC CTG CGC ACC GGA CAG GAG ATG CCG ATG GAA ATG GCG
29   I   D   Y   L   R   T   G   Q   E   M   P   M   E   M   A

136  GAT CAG GCG ATT AAC GTG GTG CCG GGC ATG ACG CCG AAA ACC ATT
44   D   Q   A   I   N   V   V   P   G   M   T   P   K   T   I

181  CTT CAC GCC GGG CCG CCG ATC CAG CCT GAC TGG CTG AAA TCG AAT
59   L   H   A   G   P   P   I   Q   P   D   W   L   K   S   N

226  GGT TTT CAT GAA ATT GAA GCG GAT GTT AAC GAT ACC AGC CTC TTG
74   G   F   H   E   I   E   A   D   V   N   D   T   S   L   L

271  CTG AGT GGA GAT GCC TCC TAC CCT TAT GAT GTG CCA GAT TAT GCC
89   L   S   G   D   A   S   Y   P   Y   D   V   P   D   Y   A

316  TCT CCC GAA TTC GGC CGA CTC GAG AAG CTT
104  S   P   E   F
```

Fig. 3C

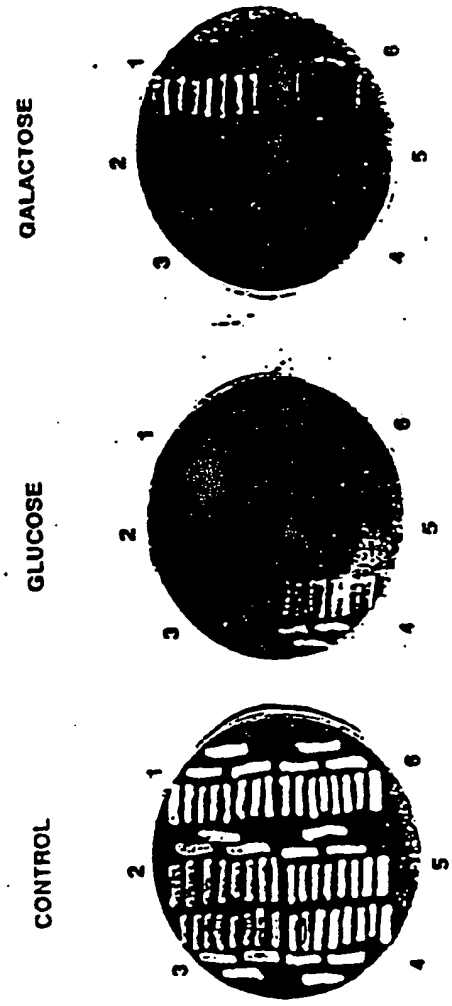


Fig. 4

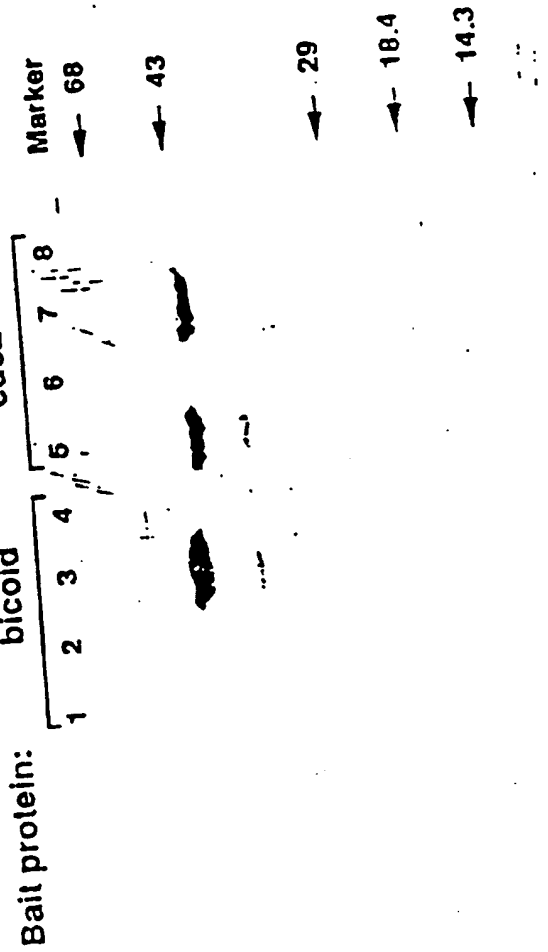


Fig. 5

```
                -31  C GGC ACT GGT CTC GAC GTG GGG CGG CCA GCG

1     ATG GAG CCG CCC AGT TCA ATA CAA ACA AGT GAG TTT GAC TCA TCA GAT
1      M   E   P   P   S   S   I   Q   T   S   E   F   D   S   S   D

49    GAA GAG CCT ATT GAA GAT GAA CAG ACT CCA ATT CAT ATA TCA TGG CTA
17     E   E   P   I   E   D   E   Q   T   P   I   H   I   S   W   L

97    TCT TTG TCA CGA GTG AAT TGT TCT CAG TTT CTC GGT TTA TGT GCT CTT
33     S   L   S   R   V   N   C   S   Q   F   L   G   L   C   A   L

145   CCA GGT TGT AAA TTT AAA GAT GTT AGA AGA AAT GTC CAA AAA GAT ACA
49     P   G   C   K   F   K   D   V   R   R   N   V   Q   K   D   T

193   GAA GAA CTA AAG AGC TGT GGT ATA CAA GAC ATA TTT GTT TTC TGC ACC
65     E   E   L   K   S   C   G   I   Q   D   I   F   V   F   C   T

241   AGA GGG GAA CTG TCA AAA TAT AGA GTC CCA AAC CTT CTG GAT CTC TAC
81     R   G   E   L   S   K   Y   R   V   P   N   L   L   D   L   Y

289   CAG CAA TGT GGA ATT ATC ACC CAT CAT CAT CCA ATC GCA GAT GGA GGG
97     Q   Q   C   G   I   I   T   H   H   H   P   I   A   D   G   G

337   ACT CCT GAC ATA GCC AGC TGC TGT GAA ATA ATG GAA GAG CTT ACA ACC
113    T   P   D   I   A   S   C   C   E   I   M   E   E   L   T   T

385   TGC CTT AAA AAT TAC CGA AAA ACC TTA ATA CAC TGC TAT GGA GGA CTT
129    C   L   K   N   Y   R   K   T   L   I   H   C   Y   G   G   L

432   GGG AGA TCT TGT CTT GTA GCT GCT TGT CTC CTA CTA TAC CTG TCT GAC
145    G   R   S   C   L   V   A   A   C   L   L   L   Y   L   S   D

481   ACA ATA TCA CCA GAG CAA GCC ATA GAC AGC CTG CGA GAC CTA AGA GGA
161    T   I   S   P   E   Q   A   I   D   S   L   R   D   L   R   G

529   TCC GGG GCA ATA CAG ACC ATC AAG  CAA TAC AAT TAT CTT CAT GAG TTT
177    S   G   A   I   Q   T   I   K   Q   Y   N   Y   L   H   E   F

577   CGG GAC AAA TTA GCT GCA CAT CTA TCA TCA AGA GAT TCA CAA TCA AGA
192    R   D   K   L   A   A   H   L   S   S   R   D   S   Q   S   R

625   TCT GTA TCA AGA TAA AGG AAT TCA AAT AGC ATA TAT ATG ACC ATG TCT
209    S   V   S   R   *

673   GAA ATG TCA GTT CTC TAG CAT AAT TTG TAT TGA AAT GAA ACC ACC AGT

721   GTT ATC AAC TTG AAT GTA AAT GTA CAT GTG CAG ATA TTC CTA AAG TTT

769   TAT TGA C
```

Fig. 6

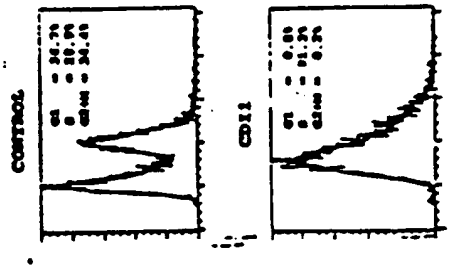


Fig. 7C

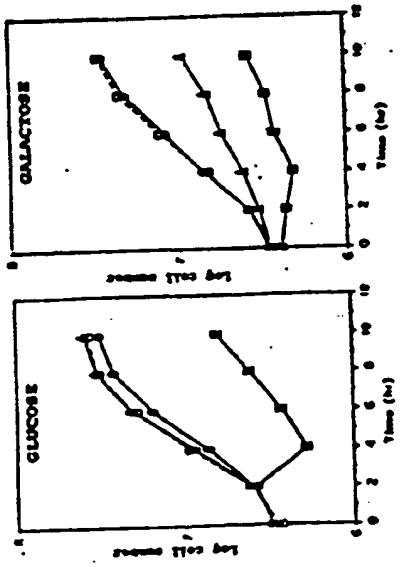


Fig. 7A

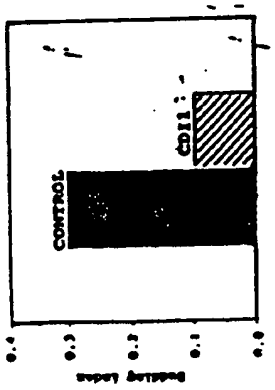


Fig. 7B

Fig. 8C

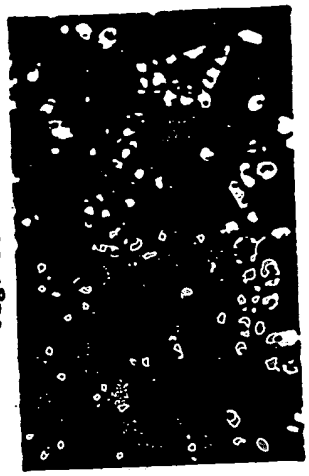


Fig. 8D

Fig. 8A

Fig. 8B

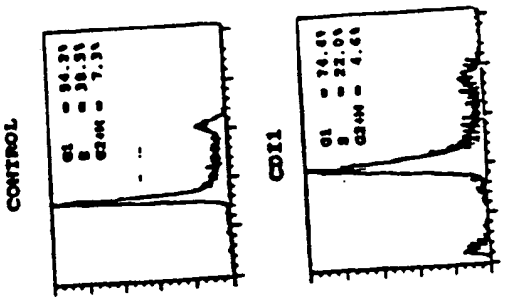


Fig. 9B

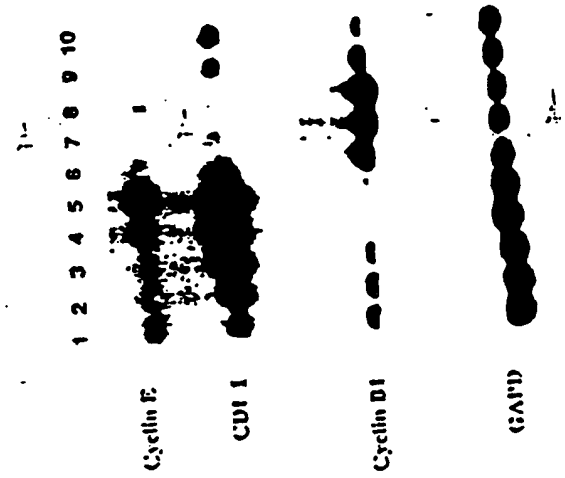


Fig. 9A

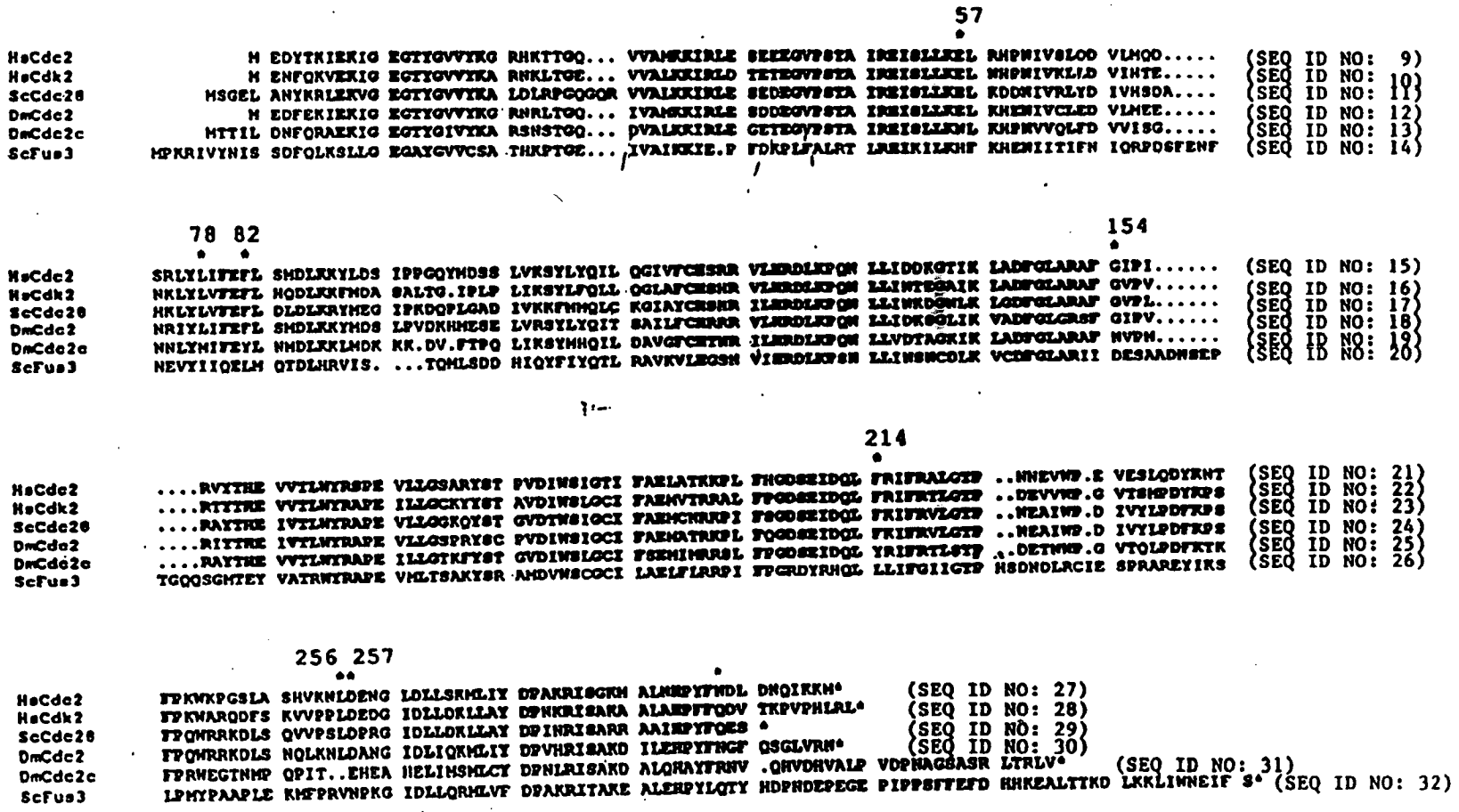

57
78 82
154
214
256 257
HsCdc2 (SEQ ID NO: 9)
HsCdk2 (SEQ ID NO: 10)
ScCdc28 (SEQ ID NO: 11)
DmCdc2 (SEQ ID NO: 12)
DmCdc2c (SEQ ID NO: 13)
ScFus3 (SEQ ID NO: 14)
HsCdc2 (SEQ ID NO: 15)
HsCdk2 (SEQ ID NO: 16)
ScCdc28 (SEQ ID NO: 17)
DmCdc2 (SEQ ID NO: 18)
DmCdc2c (SEQ ID NO: 19)
ScFus3 (SEQ ID NO: 20)
HsCdc2 (SEQ ID NO: 21)
HsCdk2 (SEQ ID NO: 22)
ScCdc28 (SEQ ID NO: 23)
DmCdc2 (SEQ ID NO: 24)
DmCdc2c (SEQ ID NO: 25)
ScFus3 (SEQ ID NO: 26)
HsCdc2 (SEQ ID NO: 27)
HsCdk2 (SEQ ID NO: 28)
ScCdc28 (SEQ ID NO: 29)
DmCdc2 (SEQ ID NO: 30)
DmCdc2c (SEQ ID NO: 31)
ScFus3 (SEQ ID NO: 32)

Fig. 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US93/10069

A. CLASSIFICATION OF SUBJECT MATTER

IPC(5)- :C12N 15/10, 1/19; C12Q 1/68
US CL :435/6, 69.1, 172.3, 320.1, 255; 530/358; 536/23.5
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 69.1, 172.3, 320.1, 255; 530/358; 536/23.5

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS, EMBASE, MEDLINE, LIFESCI, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE, Volume 89, issued September 1992, E.R. Fearon et al., "Karyoplasmic Interaction Selection Strategy: A General Strategy to Detect Protein-Protein Interactions in Mammalian Cells", pages 7958-7962, entire document.	1-10
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE, Volume 88, issued November 1991, C. Chien et al., "The Two-Hybrid System: A Method to Identify and Clone Genes For proteins That Interact With a Protein of Interest", pages 9578-9582, entire document.	1-10

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be part of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 26 November 1993	Date of mailing of the international search report JAN 26 1994
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. NOT APPLICABLE	Authorized officer ROBERT WAX Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US93/10069

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	NATURE, Volume 340, issued 20 July 1989, S. Fields et al., "A Novel Genetic System to Detect Protein-Protein Interactions", pages 245-246, entire document.	1-10
Y	JOURNAL OF NIH RESEARCH, Volume 3, No. 12, issued December 1991, Nancy Touchette, "New Approach Detects Protein Interactions In Vivo", pages 44-46, entire document.	1-10
Y	NATURE, Volume 334, issued 25 August 1988, G. Gill et al., "Negative Effect of the Transcriptional Activator GAL4", pages 721-724, entire document.	1-10
Y	CELL, Volume 70, issued 24 July 1992, S.L. Berger et al., "Genetic Isolation of ADA2: A Potential Transcriptional Adaptor Required for Function of Certain Acidic Activation Domains", pages 251-265, entire document.	1-10
Y	CELL, Volume 51, issued 09 October 1987, J. Ma et al., "A New Class of Yeast Transcriptional Activators", pages 113-119, entire document.	1-10
Y	TRENDS IN BIOCHEMICAL SCIENCE, Volume 15, issued 1990, G. Draetta, "Cell Cycle Control in Eukaryotes: Molecular Mechanisms of Cdc2 Activation", pages 378-383, entire document.	10

molecules released from the beads (while the beads are retained) enabling the compounds to diffuse into the surrounding medium. The effects, such as plaques with a bacterial lawn, can be observed. Zones of growth inhibition or growth activation or effects on gene expression can then be visualized and the beads at the center of the zone picked and analyzed.

One assay scheme will involve gels where the molecule or system, e.g. cell, to be acted upon may be embedded substantially homogeneously in the gel. Various gelling agents may be used such as polyacrylamide, agarose, gelatin, etc. The particles may then be spread over the gel so as to have sufficient separation between the particles to allow for individual detection. If the desired product is to have hydrolytic activity, a substrate is present in the gel which would provide a fluorescent product. One would then screen the gel for fluorescence and mechanically select the particles associated with the fluorescent signal.

One could have cells embedded in the gel, in effect creating a cellular lawn. The particles would be spread out as indicated above. Of course, one could place a grid over the gel defining areas of one or no particle. If cytotoxicity were the criterion, one could release the product, incubate for a sufficient time, followed by spreading a vital dye over the gel. Those cells which absorbed the dye or did not absorb the dye could then be distinguished.

As indicated above, cells can be genetically engineered so as to indicate when a signal has been transduced. There are many receptors for which the genes are known whose expression is activated. By inserting an exogenous gene into a site where the gene is under the transcriptional control of the promoter responsive to such receptor, an

enzyme can be produced which provides a detectable signal, e.g. a fluorescent signal. The particle associated with the fluorescent cell(s) may then be analyzed for its reaction history.

Libraries and Kits

For convenience, libraries and/or kits may be provided. The libraries would comprise the particles to which a library of products and tags have been added so as to allow for screening of the products bound to the bead or the libraries would comprise the products removed from the bead and grouped singly or in a set of 10 to 100 to 1000 members for screening. The kits would provide various reagents for use as tags in carrying out the library syntheses. The kits will usually have at least 4, usually 5, different compounds in separate containers, more usually at least 10, and may comprise at least 10^2 different separated organic compounds, usually not more than about 10^2, more usually not more than about 36 different compounds. For binary determinations, the mode of detection will usually be common to the compounds associated with the analysis, so that there may be a common chromophore, a common atom for detection, etc. Where each of the identifiers is pre-prepared, each will be characterized by having a distinguishable composition encoding choice and stage which can be determined by a physical measurement and including groups or all of the compounds sharing at least one common functionality.

Alternatively, the kit can provide reactants which can be combined to provide the various identifiers. In this situation, the kit will comprise a plurality of separated first functional, frequently bifunctional, organic compounds, usually four or more, generally one for each stage of the synthesis, where the functional organic compounds share the same functionality and are distinguishable as to at least one determinable

characteristic. In addition, one would have at least one, usually at least two, second organic compounds capable of reacting with a functionality of the functional organic compounds and capable of forming mixtures which are distinguishable as to the amount of each of said second organic compounds. For example, one could have a glycol, amino acid, or a glycolic acid, where the various bifunctional compounds are distinguished by the number of fluorine or chlorine atoms present, to define stage, and have an iodomethane, where one iodomethane has no radioisotope, another has ^{14}C and another has one or more ^{3}H. By using two or more of the iodomethanes, one could provide a variety of mixtures which could be determined by their radioemissions. Alternatively, one could have a plurality of second organic compounds, which could be used in a binary code.

As indicated previously one could react the tags after release with a molecule which allows for detection. In this way the tags could be quite simple, having the same functionality for linking to the particle as to the detectable moiety. For example, by being linked to a hydroxycarboxyl group, an hydroxyl group would be released, which could then be esterified or etherified with the molecule which allows for detection. For example, by using combinations of fluoro- and chloroalkyl groups, in the binary mode, the number of fluoro and/or chloro groups could determine choice, while the number of carbon atoms would indicate stage.

Groups of compounds of particular interest include linkers joined to a substituted ortho-nitrobenzyloxy group, indanyloxy or fluorenyloxy group, or other group which allows for photolytic or other selective cleavage. The linking group may be an alkylene group of from 2 to 20 carbon atoms, polyalkyleneoxy, particularly alkyleneoxy of from 2 to 3 carbon atoms, cycloalkyl group of from 4 to 8

carbon atoms, haloalkyl group, particularly fluoroalkyl of from 2 to 20 carbon atoms, one or more aromatic rings and the like, where the linker provides for the discrimination between the various groups, by having different numbers of units and/or substituents.

Individual particles or a plurality of particles could be provided as articles of commerce, particularly where the particle(s) have shown a characteristic of interest. Based on the associated tags, the reaction history may be decoded. The product may then be produced in a large synthesis. Where the reaction history unequivocally defines the structure, the same or analogous reaction series may be used to produce the product in a large batch. Where the reaction history does not unambiguously define the structure, one would repeat the reaction history in a large batch and use the resulting product for structural analysis. In some instances it may be found that the reaction series of the combinatorial chemistry may not be the preferred way to produce the product in large amounts.

Thus, an embodiment of this invention is a kit comprising a plurality of separated organic compounds, each of the compounds characterized by having a distinguishable composition, encoding at least one bit of different information which can be determined by a physical measurement, and sharing at least one common functionality. A preferred embodiment is a kit comprising at least 4 different functional organic compounds.

More preferred is a kit wherein said functional organic compounds are of the formula:

$$F^1\text{-}F^2\text{-}\underline{C}\text{-}E\text{-}\underline{C}' \qquad I$$

where F^1-F^2 is a linker which allows for attachment to and detachment from a solid particle; and

C-E-C' is a tag member which can be determined by a physical measurement, especially wherein said functional organic compounds differ by the number of methylene groups and/or halogens, nitrogens or sulfurs present.

Also preferred is a kit wherein the C-E-C' portion is removed photochemically or a kit wherein the C-E-C' portion is removed oxidatively, hydrolytically, thermolytically, or reductively.

Compounds of this invention may be useful as analgesics and/or for the treatment of inflammatory disease, especially in the case of the azotricyclics acting as antagonists of the meurokin 1/brandykin receptor. Members of the benzodiazopine library may be useful as a muscle relaxant and/or tranquilizer and/or as a sedative. Members of the 23 million Mixed Amide Library may be of use in the treatment of hypertension on endothelin antagonists or Raynaud's syndrcme.

The following examples are offered by way of illustration and not by way limitation.

In one embodiment the invention is composition comprising at least 6 different components, each component having a distinguishable moiety. The components may be characterized by each moiety being substantially chemically stable or inert and having an identifiable characteristic different from each of the other moieties. Each moiety is joined to a linking group having an active functionality capable of forming a covalent bond, through a linking group to individually separable solid surfaces, or joined to a group which is detectable at less than 1 nanomole. With a proviso that when the moieties are joined to the linking group, the components are physically segregated. Preferably, the solid supports are beads.

In one embodiment each component comprises molecules of different compounds bound to individual separable solid surfaces, wherein the molecules on the solid surfaces. Preferably, the moieties of the invention define an homologous series and/or a series of substitutions on a core molecule.

The invention herein is also directed to a compound library comprising at least one hundred unique solid supports. In this compound library each solid support has (1) an individual compound bound to the solid support as a major compound bound to the support; and (2) a plurality of tags e.g. tags incapable of being sequenced, where the tags are individual tag molecules which are physically distinguishable in being physically separable and are substituted so as to be detectable at less than about a nanomole or have a functional group for bonding to a substituent which is detectable at less than about at nanomole. Preferably, in the compound library each solid support has at least about 6 tags. In another embodiment, in the compound library the tags define a binary code encoding the synthetic protocol used for the synthesizing the compound on the solid support.

This invention also provides a method for determining a synthetic protocol encoded by separable physically different tags in a series and defining a binary code. In this method at least two tags are employed to define each stage of the synthetic protocol, there being at least six tags. The step of the method comprising separating tags by means of their physical differences and detecting the tags to define a binary line encoding the protocol whereby the synthetic protocol is determined in accordance with the binary line.

Compound of this invention may be useful as analgesics and/or for the treatment of inflammatory disease,

especially in the case of the azatricyclics acting as antagonists of the neurokinin 1/brandykin receptor. Members of the benzodiazopine library may be useful as a muscle relaxant and/or tranquilizer and/or as a sedative. Members of the 23 Mixed Amide Library may be of use in the treatment of hypertension on endothelin antagonists or Raynaud's syndrome.

EXAMPLE 1

PEPTIDE LIBRARY

In order to encode up to 10^9 different syntheses, one could prepare 30 different identifiers which carry individual tags capable of being separated one from another by capillary GC. For encoding a smaller number of syntheses, fewer identifiers would be used. The tags would normally be prepared from commercially-available chemicals as evidenced by the following illustration. ω-Hydroxyalkenes-1, where the number of methylene groups would vary from 1 to 5, would be reacted with an iodoperfluoroalkane, where the number of CF_2 groups would be 3, 4, 6, 8, 10, and 12. By employing a free-radical catalyst, the iodoperfluorocarbon would add to the double bond, where the iodo group could then be reduced with hydrogen and a catalyst or a tin hydride. In this manner, 30 different tags could be prepared. The chemical procedure is described by Haszeldine and Steele, J. Chem. Soc. (London), 1199 (1953); Brace, J. Fluor. Chem., 20, 313 (1982). The highly fluorinated tags can be easily detected by electron capture, have different GC retention times, so that they are readily separated by capillary GC, are chemically inert due to their fluorinated, hydrocarbon structure and each bears a single hydroxyl functional group for direct or indirect attachment to particles.

Before attachment to compound precursors, the tags (referred to as T1-T30) would be activated in a way which is appropriate for the chemical intermediates to be used

in the combinatorial synthesis. By appropriate it is intended that a functionality would be added which allows for ready attachment by a chemical bond to a compound precursor or to the bead matrix itself. The activation process would be applied to each of the 30 different tags and allow these tags to be chemically bound, either directly or indirectly, to intermediates in the combinatorial compound synthesis. For example, a carboxy derivative could be used for coupling and upon activation the resulting carboxy group would bond to the particle.

In the case of a combinatorial synthesis of a peptidic compound or other structure made of amide-linked organic fragments, the encoding process could consist of addition of a carboxylic acid-equipped linker. For example, the tag would be coupled to the tert.-butyl ester of o-nitro-p-carboxybenzyl bromide in the presence of sodium hydride. The ester would then be hydrolyzed in dilute trifluoroacetic acid.

Activated identifiers would be coupled to intermediates at each stage in the combinatorial compound synthesis. The ortho-nitrobenzyl ether part of the activated identifiers is used to allow photochemical detachment of the tags after completing the combinatorial synthesis and selecting the most desirable compounds. The detached tags would then be decoded using capillary GC with electron capture detection to yield a history of the synthetic stages used to prepare the compound selected.

While there is an almost unlimited set of chemical stages and methods which could be used to prepare combinatorial libraries of compounds, we will use coupling of α-amino acids to make a combinatorial library of peptides as an example of an application of the encoding methodology. In this example, we will describe preparation of a library of pentapeptides having all combinations of 16 different

amino acids at each of the five residue positions. Such a library would contain 16^5 members. To uniquely encode all members of this library, 20 detachable tags (T1-T20) as described above would be required.

To prepare the encoded library, we would begin with a large number ($>10^6$) of polymer beads of the type used for Merrifield solid phase synthesis and functionalized by free amino groups. We would divide the beads into 16 equal portions and place a portion in each of 16 different reaction vessels (one vessel for each different α-amino acid to be added). We would then add a small portion (e.g., 1 mol%) of identifiers to each of the amino acid derivatives (e.g., Fmoc amino acids) to be coupled in the first stage of the combinatorial synthesis. The specific combination of the tags incorporated into the identifiers added would represent a simple binary code which identifies the amino acid used in the first stage of synthesis. The 16 amino acids added would be indicated by numbers 1-16 and any such number could be represented chemically by combinations of the first four tags (T1-T4). In tables 2 and 3, a typical encoding scheme is shown in which the presence or absence of a tag is indicated by a 1 or a 0, respectively. The letter T may represent either the the tag or the identifier incorporating that tag.

Table 2. A typical encoding scheme.

Amino Acid added in first stage	T4	T3	T2	T1
Number 1 (e.g., glycine	0	0	0	0
Number 2 (e.g., alanine)	0	0	0	1
Number 3 (e.g., valine)	0	0	1	0
Number 4 (e.g., serine)	0	0	1	1
Number 5 (e.g., threonine)	0	1	0	0
.				
.				
Number 16 (e.g., tryptophan)	1	1	1	1

We would then carry out a standard dicyclohexyl-carbodiimide (DCC) peptide coupling in each of the 16 vessels using the Fmoc amino acids admixed with small amounts of the encoding activated identifiers as indicated above. During the couplings, the amino acids as well as small amounts (e.g., 1%) of the identifiers would become chemically bound to intermediates attached to the beads.

Next the beads would be thoroughly mixed and again separated into 16 portions. Each portion would again be placed in a different reaction vessel. A second amino acid admixed with appropriate new activated identifiers (T5-T8) would be added to each vessel and DCC coupling would be carried out as before. The particular mixture of the incorporated tags (T5-T8) would again represent a

simple binary code for the amino acid added in this, the second stage of the combinatorial synthesis.

Table 3. A typical encoding scheme.

Amino Acid added in second stage	T8	T7	T6	T5
Number 1 (e.g., glycine	0	0	0	0
Number 2 (e.g., alanine)	0	0	0	1
Number 3 (e.g., valine)	0	0	1	0
Number 4 (e.g., serine)	0	0	1	1
Number 5 (e.g., threonine)	0	1	0	0
.				
.				
Number 16 (e.g., tryptophan)	1	1	1	1

After the 16 couplings of stage 2 are complete, the beads would be again mixed and then divided into 16 new portions for the third stage of the synthesis. For the third stage, T9-T12 would be used to encode the third amino acid bound to the beads using the same scheme used for stages 1 and 2. After the third couplings, the procedure would be repeated two more times using the fourth amino acids with T13-T16 and the fifth amino acids with T17-T20 to give the entire library of 1,048,576 different peptides bound to beads.

Although the above beads would be visually indistinguishable, any bead may be chosen (e.g., by

selecting based on the interesting chemical or biological properties of its bound peptide or other target molecule) and its synthetic history may be learned by detaching and decoding the associated tags.

The precise method used to detach tags will depend upon the particular linker used to chemically bind it to intermediates in the combinatorial synthesis of the target compound. In the example above, the ortho-nitrobenzyl carbonate linkages, which are known to be unstable to ~300 nm light (Ohtsuka, et al., J. Am. Chem. Soc., 100, 8210 [1978]), would be cleaved by photochemical irradiation of the beads. The tags would then diffuse from the beads into free solution which would be injected into a capillary gas chromatograph (GC) equipped with a sensitive electron capture detector. Since the order in which the tags (T1-T20) emerged from the GC and their retention times under standard conditions were previously determined, the presence or absence of any of T1-T20 would be directly determined by the presence or absence of their peaks in the GC chromatogram. If 1 and 0 represent the presence and absence respectively of peaks corresponding to T1-T20, then the chromatogram can be taken as a 20-digit binary number which can uniquely represent each possible synthesis leading to each member of the peptide library. The use of halocarbon tags which are safe, economical and detectable at subpicomole levels by

electron capture detection makes this capillary GC method a particularly convenient encoding scheme for the purpose.

As an example of using the encoding scheme for the pentapeptide library above, a particular bead is irradiated with light to detach the tags, the solubilized labels injected into a capillary GC and the following chromatogram obtained ("Peak" line):

Label	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	GC Inject
Peak	|	|	|	|		|					|	|			|			|			|
Binary	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	
Stage	----5-----				-----4-----				----3------				-----2-----				-----1-----				
AA	Tryptophan				Threonine				Serine				Alanine				Valine				

The "Label" line diagrams the GC chromatogram where T20-T1 peaks (|) are to be found (note the injection is given on the right and the chromatogram reads from right to left). The "Peak" line represents the presence of labels (T20-T1) as peaks in the chromatogram. The "Binary" line gives presence (1) or absence (0) of peaks as a binary number. The "Stage" line breaks up the binary number into the five different parts encoding the five different stages in the synthesis. Finally, the "AA" line gives the identity of the amino acid which was added in each stage and was given by the binary code in the "Binary" line above.

EXAMPLE 2

RADIO-LABELED TAGS

In the next illustration, the tags employed are monomethylethers of linear alkyl-α,ω-diols. The diol would have N + 2 carbon atoms, where N designates the stage. The methyl group would be a radiolabeled reagent which would have any of a variety of $^{3}H/^{14}C$ ratios from 1/1 to m/1, where m is the number of choices. The double radiolabel allows for accurate quantitation of the tritium present in the tag. By having 10 different alkylene groups and 10 different radioactive label ratios, 10^{10} unique ten-member sets of tags are generated. Tags would be attached by first reacting them with activating agents, e.g. phosgene to form a chloroformate, followed by reaction with the F^1-F^2 component. In this case, F^1-F^2 is the o-nitro-p-carboxy-benzyl alcohol protected as the t-butyl ester. Each time a synthetic stage is carried out, the de-esterified identifier is added directly to the bead, which has covalently bonded amine or hydroxyl groups, to form amides or esters with the acid activated using standard chemistry, e.g., carbodiimide coupling methodology. At the end of the sequential synthesis, the beads are then screened with a variety of receptors or enzymes to determine a particular characteristic. The beads demonstrating the characteristic may then be isolated, the tags detached and separated by HPLC to give a series of glycol monomethyl ethers which may then be analyzed for radioactivity by standard radioisotope

identification methods. For example, if the first and second tags to elute from the HPLC column had $^{3}H/^{14}C$ ratios of 5:1 and 7:1 respectively, then the product which showed activity had been synthesized by reagent number 5 in stage 1 and reagent number 7 in stage 2.

EXAMPLE 3

2401 Peptide Library

The identifiers employed were 2-nitro-4-carboxybenzyl, O-aryl substituted ω-hydroxyalkyl carbonate, where alkyl was of from three to 12 carbon atoms and aryl was (A) pentachlorophenyl, (B) 2,4,6-trichlorophenyl, or (C) 2,6-dichloro-4-fluorophenyl. The tags are designated as NAr, wherein N is the number of methylene groups minus two and Ar is the aryl group. Thus, tag 2A has a butylene group bonded to the pentachlorophenyl through oxygen. The subject tags can be easily detected using electron capture gas chromatography at about 100 fmol.

In the subject analysis, the tagging molecules are arranged in their GC elution order. Thus the tag which is retained the longest on the GC column is designated T1 and is associated with the least significant bit in the binary synthesis code number, the next longest retained tag is called T2 representing the next least significant binary bit, and so on. Using an 0.2mM x 20M methylsilicone capillary GC column, eighteen well-resolved tags were obtained where T1 through T18 corresponded to 10A, 9A, 8A,

7A, 6A, 5A, 4A, 3A, 6B, 2A, 5B, 1A, 4B, 3B, 2B, 1B, 2C, and 1C, respectively.

An encoded combinatorial library of 2401 peptides was prepared. This library had the amino acid sequence N-XXXXEEDLGGGG-bead, where the variable X residues were D, E, I, K, L, Q, or S (single letter code). The 4 glycines served as a spacer between the encoded amino acid sequence and the bead. The combinatorial library included the sequence H_2N-KLISEEDL, part of the 10 amino acid epitope which is known to be bound by 9E10, a monoclonal antibody directed against the human C-myc gene product. For encoding this library, three binary bits were sufficient to represent the seven alternative reagents for each stage. The code was as follows: 001 = S; 010 = I; 011 = K; 100 = L; 101 = Q; 110 = E; 111 = D.

The library was synthesized by first preparing the constant segment of the library H_2NEEDLGGGG-bead on 1.5 g of 50-90μ polystyrene synthesis beads functionalized with 1.1 meq/g of aminomethyl groups using standard solid phase methods based on t.-butyl side-chain protection and Fmoc main chain protection (Stewart and Young, "Solid Phase Peptide Synthesis", 2nd edition, Pierce Chemical Co., 1984). After deprotecting the Fmoc groups with diethylamine, the beads were divided into seven 200 mg fractions and each fraction placed in a different Merrifield synthesis vessel mounted on a single wrist-

action shaker. The beads in the seven vessels were processed independently as follows (see Table 3-1). The letter T in this example refers to the tag or to the identifier incorporating that tag.

TABLE 3-1

Vessel No.	Step 1	Step 2	Step 3	Step 4
1	1%T1	DIC, wash	Fmoc(tBu)S, Anh.	Wash
2	1%T2	"	FmocI, Anh.	"
3	1%T1,T2	"	Fmoc(Boc)K, Anh.	"
4	1%T3	"	FmocL, Anh.	"
5	1%T1,T3	"	Fmoc(trityl)Q, Anh.	"
6	1%T2,T3	"	Fmoc(t-butyl)E, Anh.	"
7	1%T1,T2,T3	"	Fmoc(tBu)D, Anh.	"

In accordance with the above procedure a sufficient amount of the identifiers listed in step 1 were attached via their carboxylic acids using diisopropylcarbodiimide to tag about 1% of the free amino groups on each bead in the corresponding vessel. The remaining free amino groups on each bead were then coupled in step 3 to N-protected amino acid anhydrides. After washing with methylene chloride, isopropanol, and N,N-dimethylformamide, the beads from the seven vessels were combined and thoroughly mixed. At this point the library had seven members.

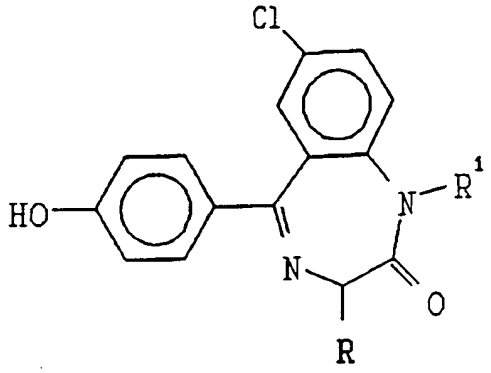
After Fmoc deprotection (diethylamine), the beads were again divided into seven vessels and processed as before except that in place of the identifiers used previously, identifiers representing the second stage (T4-6) were used. By repeating the procedure two more times, using identifiers T7-9 and then T10-12 analogously, the entire uniquely encoded library of 7^4=2401 different peptides was prepared using only 12 identifiers.

To read the synthesis code from a single selected bead, the bead was first washed four times in a small centrifuge tube with 100 μL portions of DMF, and then resuspended in 1 μL of DMF in a Pyrex capillary tube. After 2 hrs of photolysis with a Rayonet 350 nm light source, the tags released from the bound identifiers were silylated using about 0.1 μL bis-trimethylsilylacetamide and the solution injected into a Hewlett Packard capillary gas chromatograph equipped with an 0.2mM x 20M methylsilicone fused silica capillary column and an electron capture detector. The binary synthesis code of the selected bead was directly determined from the chromatogram of the tags which resulted.

EXAMPLE 4

Benzodiazepine Library

A combinatorial benzodiazepine library comprising 30 compounds of the formula VIII



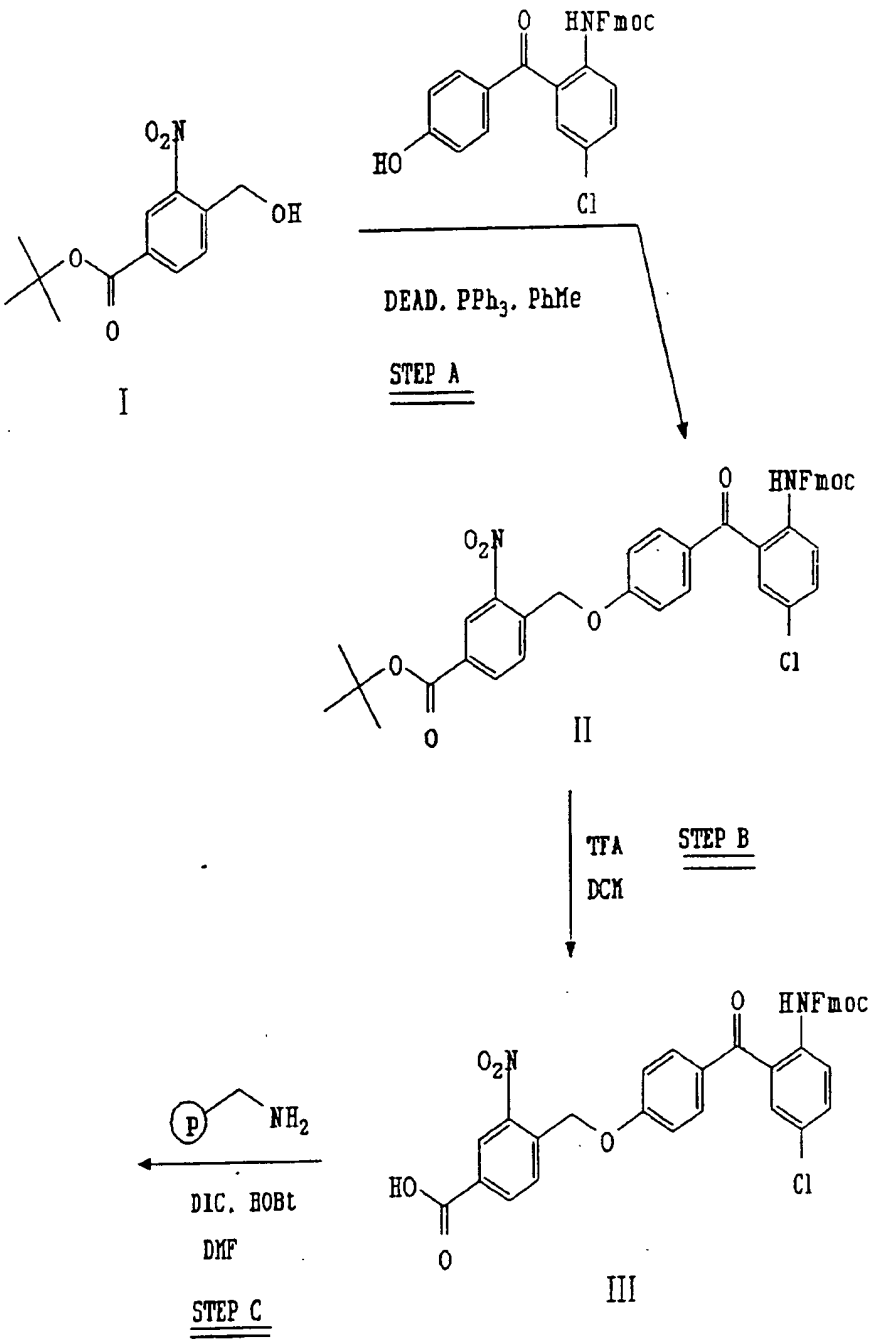
VIII

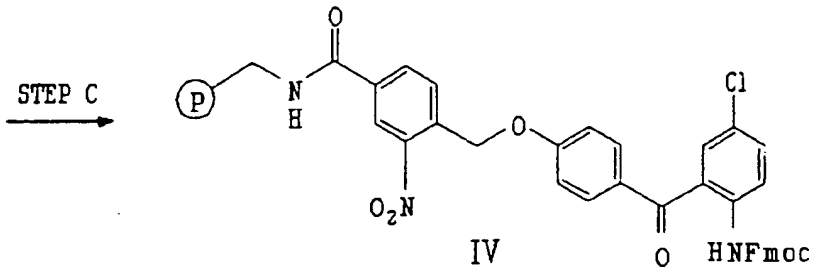
wherein:

R is CH_3, $CH(CH_3)_2$, CH_2CO_2H, $(CH_2)_4NH_2$, $CH_2C_6H_4OH$, or $CH_2C_6H_5$ and

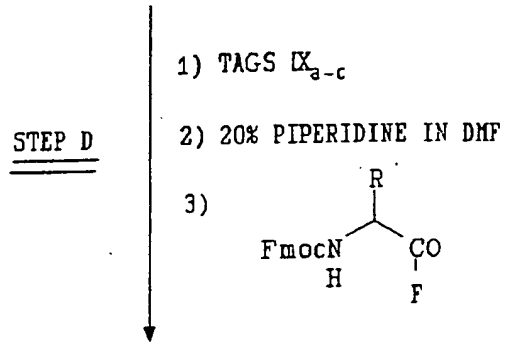
R^1 is H, CH_3, C_2H_5, $CH_2CH{=}CH_2$, or $CH_2C_6H_5$

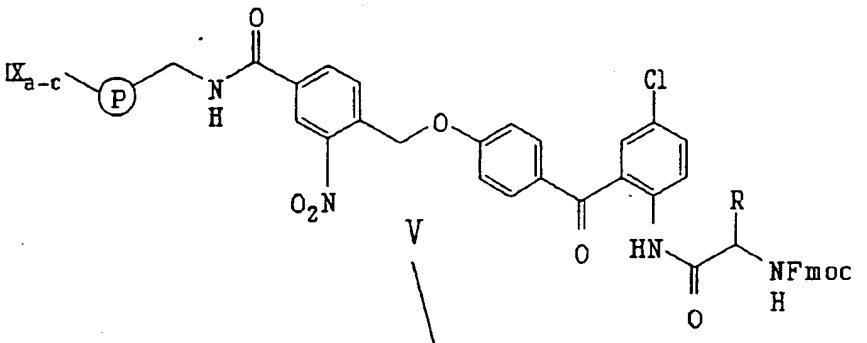
is constructed per the following scheme.

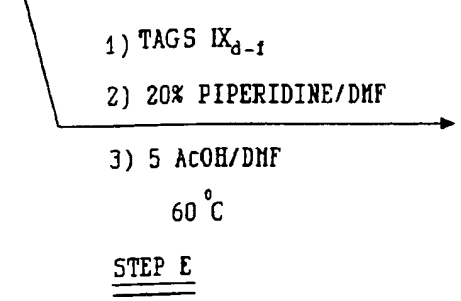

HNFmoc
HO
Cl
O_2N
OH
O
O
O
I
DEAD. PPh_3. PhMe
STEP A
HNFmoc
O_2N
O
Cl
O
O
O
II
TFA
DCM
STEP B
HNFmoc
O_2N
O
Cl
HO
O
O
III
P
NH_2
DIC. HOBt
DMF
STEP C



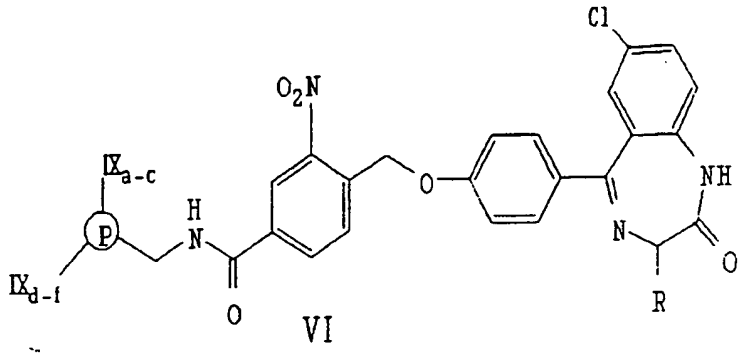
P = POLYSTYRENE RESIN







STEP E



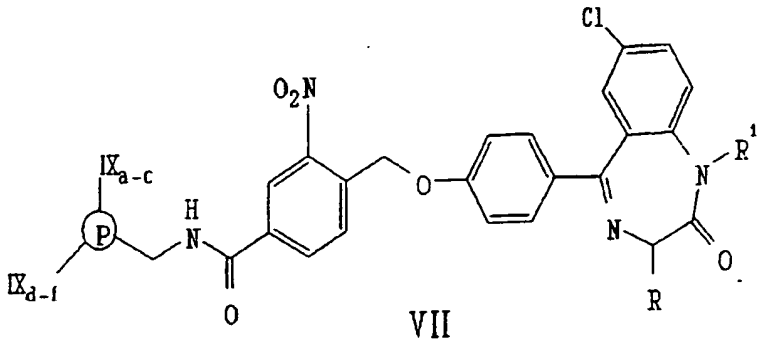
VI

STEP F

1) LITHIATED 5(PHENYLMETHYL)-2-OXAZOLIDINONE
THF, -78 °C

2) R^1X, DMF
X=BROMINE OR IODINE

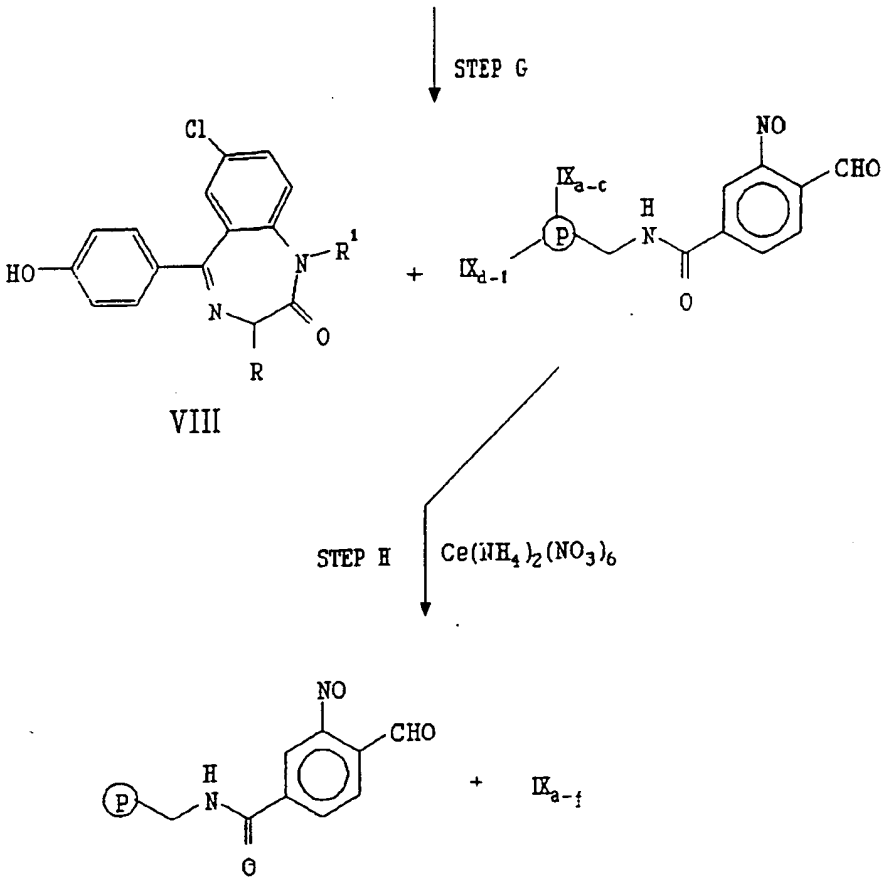
3) TFA:H_2O:DIMETHYLSULFIDE
95:5:10



VII

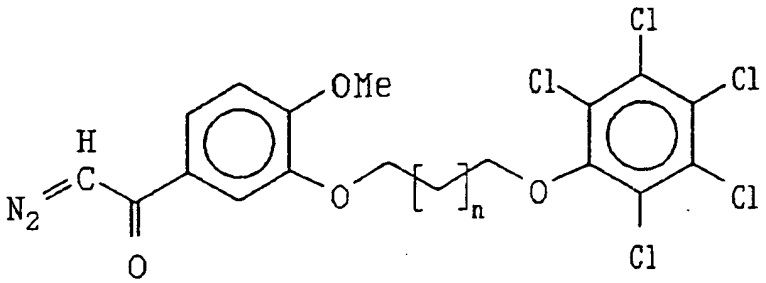
STEP G

hv (350 nm)
DMF


STEP G
Cl
HO
N
R1
N
O
R
VIII
+
IXa-c
IXd-f
P
H
N
NO
CHO
O
STEP H
Ce(NH4)2(NO3)6
NO
CHO
H
N
P
O
+
IXa-f

The benzodiazepines VIII are constructed on polystyrene beads similarly to the method of Bunin and Ellman (JACS, 114, 10997-10998 [1992]) except that a photolabile linker is incorporated between the bead and the benzodiazepine (see steps A, B, and C), thus allowing the benzodiazepine to be removed in step G non-hydrolytically by exposure to U.V. light (350 nm in DMF for 10 minutes to 12 hr). Additionally, binary codes are introduced in steps D and E which allow for a precise determination of the reaction sequence used to introduce each of the 6 R's and 5 R^1's. After removal of the tags according to step H and analysis by electron capture detection following GC separation, the nature of the individual R and R^1 groups is determined.

Steps D, E, and F essentially follow the procedure of Bunin and Ellman, but also include the incorporation of identifiers IXa-c in step D and IXd-f in Step E. The identifiers are all represented by Formula IX,



IX

wherein:

IX_a indicates n=6;

IX_b indicates n=5;

IX_c indicates n=4;

IX_d indicates n=3;

IX_e indicates n=2; and

IX_f indicates n=1.

The codes for each of R and R^1 are as follows:

Table 4-1

IX	R
a	CH_3
b	$CH(CH_3)_2$
a,b	CH_2CO_2H
c	$(CH_2)_4NH_2$
a,c	$CH_2-C_6H_4-4-OH$
b,c	$CH_2C_6H_5$
IX	R^1
d	H
e	CH_3
d,e	C_2H_5
f	$CH_2CH=CH_2$
d,f	$CH_2C_6H_5$

Step A

To a solution of I (1 equiv) in toluene (conc. = 0.5 M) is added the Fmoc protected 2-amino-5-chloro-4'-hydroxy-benzophenone (1.3 eq) and diethylazaodicarboxylate (1.3 eq) and triphenylphosphine (1.3 eq). The mixture is stirred at room temperature for 24 hr. The solvent is removed in vacuo and the residue triturated with ether and filtered and the solvent again removed in vacuo. The resultant product II is purified by chromatography on silica gel.

Step B

To a solution of II in DCM (0.2 M) stirring at r.t. is added TFA (3 equiv.) and the solution is allowed to stir for 12 hr. The solution is evaporated to dryness in vacuo and the residue dissolved in DCM, washed once with brine and dried (Na_2SO_4). Filtration and evaporation of the solvent affords III.

Step C

1% DVB (divinylbenzene) cross-linked polystyrene beads (50μ) functionalized with aminomethyl groups (1.1 mEq/g) are suspended in DMF in a peptide reaction vessel (Merrifield vessel). III (2 equiv) and HOBt (3 equiv) in DMF are added and the vessel shaken for 10 min. DIC (3 eq) is added and the vessel is shaken until a negative Ninhydrin test indicates completion of the reaction after 12 hr.

The DMF is removed and the resin washed with additional DMF (x5) and DCM (x5) before drying in vacuo.

Step D

The dry resin is divided into 6 reaction vessels and is suspended in DCM. The appropriate combinations of identifiers IX_{a-c} (see Table 4-1) are added to the flasks and shaken for 1 hr. The $Rh(TFA)_2$ catalyst (1 mol%) is added to each flask and shaken for an additional 2 hr. The flasks are drained and the resin washed with DCM (x5). The resin is then treated with a solution of TFA in DCM (0.01 M) and shaken for 30 min. and then washed again with DCM (x3) followed by DMF (x2). The resin is treated with a 20% solution of piperidine in DMF and shaken for 30 min. and is then washed with DMF (x3) and DCM (x3).

To each flask is added the appropriate Fmoc protected amino acylfluoride (3 equiv) (when required side-chain functional groups are protected as tert-butyl ester (Asp), tert-butyl ether (Tyr) or tert-butyloxycarbonyl (Lys)) with 2,6-di-tert-butyl-4-methylpyridine (10 equiv) and the flasks shaken overnight or until a negative Ninhydrin test is achieved. The resin is washed once (DCM) and then the six batches are combined and washed again (DCM, x5) before drying in vacuo.

Step E

The dry resin is divided into five reaction vessels and is suspended in DCM. The appropriate combinations of identifiers IX_{d-f} (see Table 4-1) are added to the flasks and shaken for 1 hr. The $Rh(TFA)_2$ catalyst (1 mol%) is added to each flask and shaken for an additional 2 hr. The flasks are drained and the resin washed with DCM (x5). The resin in then treated with a solution of TFA in DCM (0.01 M) and shaken for 30 min. and is then washed with DMF (x3) and DCM (x3).

To each flask is added a solution of 5% acetic acid in DMF and the mixtures are heated to 60°C and shaken overnight. The solvent is drained and then the resin washed with DMF (x5).

Step F

Each batch of resin is suspended in THF and the flasks are cooled to -78°C. To each flask is added a solution of lithiated 5-(phenylmethyl)-2-oxazolidinone (2 equiv) in THF and the mixtures are shaken at -78°C for 1 hr. The appropriate alkylating agent (Table 4-2) (4 equiv) is then added to each reaction flask followed by a catalytic amount of DMF. The vessels are allowed to warm to ambient temperature and shaken at this temperature for 5 hrs. The solvent is removed by filtration and the resin washed with THF (x1) and then dried in vacuo. The batches of resin are then combined and washed with THF (x2) and DCM (x2)

and the combined resin is then treated with a 95:5:10 mixture of TFA:water:dimethylsulphide for 2 hrs to remove the side chain protecting groups.

TABLE 4-2

IDENTIFIER	ALKYLATING AGENT
e	H_3CI
d,e	C_2H_5Br
f	$BrCH_2-CH=CH_2$
d,f	$BrCH_2C_6H_5$

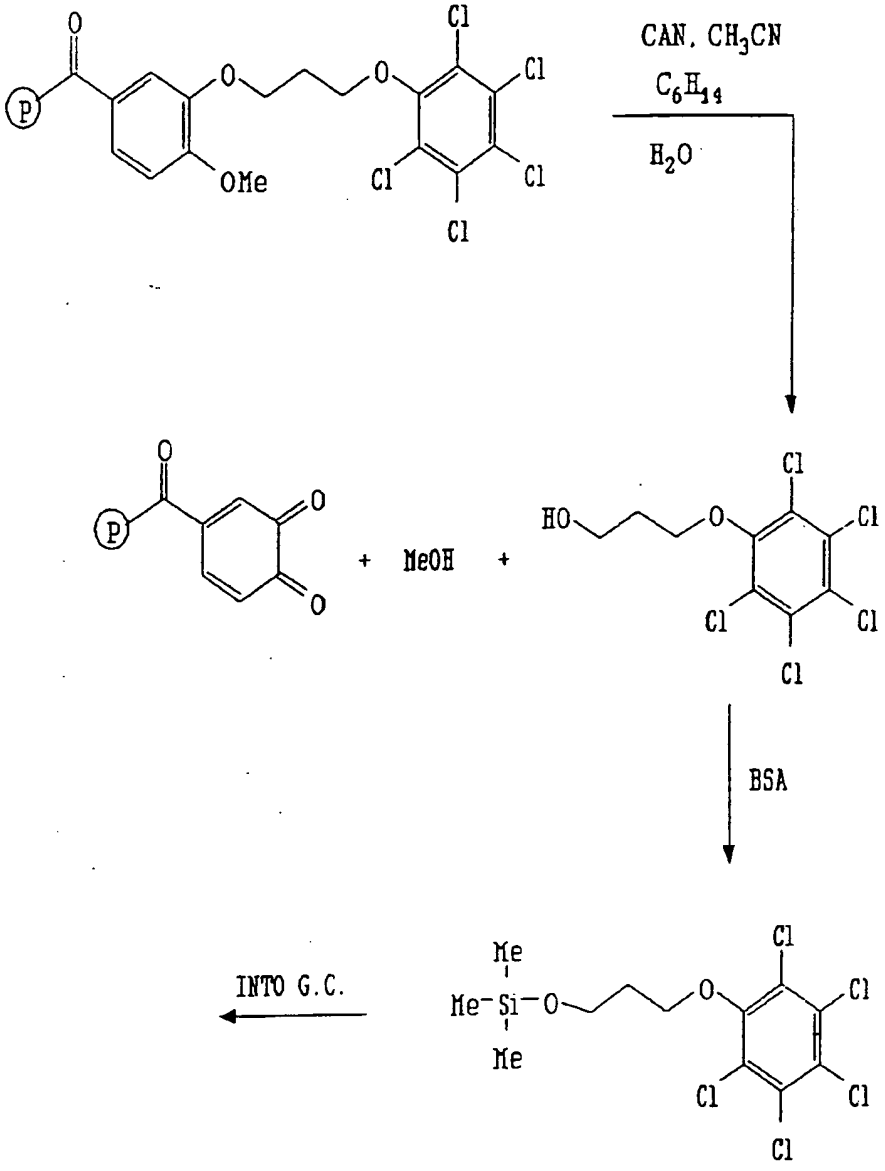
Step G

The resultant benzodiazepine can be cleaved from a bead of polystyrene by suspending the bead in DMF and irradiating with U.V. (350 nm) for 12 hrs.

Step H

A bead of interest is placed into a glass capillary tube. Into the tube is syringed 1 μL of 1M aqueous cerium (IV) ammonium nitrate (CAN) solution, 1 μL of acetonitrile and 2μL of hexane. The tube is flame sealed and then centrifuged to ensure that the bead is immersed in the reagents. The tube is placed in an ultrasonic bath and sonicated from 1 to 10 hrs preferably from 2 to 6 hrs.

The tube is cracked open and ≃1 μL of the upper hexane layer is mixed with ≃0.2 μL of bis(trimethylsilyl)-acetamide (BSA) prior to injection into the GC and each tag member determined using electron capture detection, as exemplified in the following scheme.


CAN, CH_3CN
C_6H_{14}
H_2O
MeOH
BSA
INTO G.C.
OMe
HO
Me
Si
Cl
O

EXAMPLE 5

117,649 Peptide Library

An encoded library of 117,649 peptides was prepared. This library had the sequence H_2N-XXXXXXEEDLGGGG-bead, where the variable residue X was D,E,I,K,L,Q or S. This library was encoded using the 18 tags as defined in Example 3; three binary bits being sufficient to represent the seven amino acids used in each step. The code was: 001=S; 010=I; 011=K; 100=L; 101=Q; 110=E; and 111=D, where 1 indicates the presence and 0 indicates the absence of a tag.

The constant segment of the library (H_2NEEDLGGGG-bead) was synthesized on 1.5 g of 50-80 μ Merrifield polystyrene synthesis beads functionalized with 1.1 mEq/g of aminomethyl groups using standard solid phase methods based on t-Bu sidechain protection and Fmoc mainchain protection. After deprotecting the N-terminal Fmoc protecting group with diethylamine, the beads were divided into seven 200 mg portions, each portion being placed into a different Merrifield synthesis vessel mounted on a single wrist-action shaker.

The beads in the seven vessels were processed as in Table 3-1 to attach the sets of identifiers (T1-T3) and the corresponding amino acid to each portion except that instead of DIC, i-butylchloroformate was used for activation.

This procedure first chemically attached small amounts of appropriate identifiers via their carboxylic acids to the synthesis beads. This attachment was achieved by activating the linker carboxyl groups as mixed carbonic anhydrides using isobutylchloroformate, and then adding an amount of activated identifier corresponding to 1% of the free amino groups attached to the beads. Thus, about 1% of the free amino groups were terminated for each identifier added. The remaining free amino groups were then coupled in the usual way with the corresponding protected amino acids activated as their symmetrical anhydrides.

After washing, the seven portions were combined and the Fmoc protected amino groups were deprotected by treatment with diethylamine. The beads were again divided into seven portions and processed as before, except that the appropriate identifiers carrying tags T4, T5, and T6 were added to the reaction vessels.

The procedure of dividing, labelling, coupling the amino acid combining and main-chain deprotection was carried out a total of six times using identifiers bearing tags T1-T18, affording an encoded peptide library of 117,649 different members.

Typical Identifier Preparation

To a solution of 8-bromo-1-octanol (0.91 g, 4.35 mmol) and 2,4,6-trichlorophenol (1.03 g, 5.22 mmol) in DMF (5 mL) was added cesium carbonate (1.70 g, 5.22 mmol) resulting in the evolution of gas and the precipitation of a white solid. The reaction was stirred at 80° C for 2 hrs. The mixture was diluted with toluene (50 mL) and poured into a separatory funnel, washed with 0.5 N NaOH (2x50 mL), 1N HCl (2x50 mL) and water (50 mL) and the organic phase was dried ($MgSO_4$). Removal of the solvent by evaporation gave 1.24 g (87% yield) of tag as a clear oil.

The above tag (0.81 g, 2.5 mmol) was added to a 2 M solution of phosgene in toluene (15 mL) and stirred at room temperature for 1 hr. The excess phosgene and the toluene were removed by evaporation and the resulting crude chloroformate was dissolved in DCM (5 mL) and pyridine (0.61 mL, 7.5 mmol). tert-Butyl 4-hydroxymethyl-3-nitrobenzoate (Barany and Albericio, J. Am. Chem. Soc., (1985), 107, 4936-4942) (0.5 g, 1.98 mmol) was added and the reaction mixture stirred at room temperature for 3 hrs. The solution was diluted with ethyl acetate (75 mL) and poured into a separatory funnel. After washing with 1N HCl (3x35 mL), saturated $NaHCO_3$ (2x35 mL) and brine (35 mL), the organic phase was dried ($MgSO_4$). The solvent was removed by evaporation and the residue purified by chromatography on silica gel (5% to 7.5% ethyl acetate in

petroleum ether) affording 0.95 g (79% yield) of the identifier tert-butyl ester as a clear oil.

Trifluoroacetic acid (3 mL) was added to a solution of the identifier tert-butyl ester (0.95 g, 1.57 mmol) in DCM (30 mL) to deprotect the linker acid (i.e., F^1-F^2 of Formula I) and the solution was stirred at room temperature for 7 hrs. The mixture was then evaporated to dryness and the residue redissolved in DCM (30 mL). The solution was washed with brine (20 mL) and the organic phase dried ($MgSO_4$). Removal of the solvent by evaporation gave 0.75 g (87% yield) of the identifier (6B) as a pale yellow solid. (Tag nomenclature is the same as in Example 3).

Typical Encoded Library Synthesis Step

Nα-Fmoc-E(tBu)-E(tBu)-D(tBu)-L-G4-NH-resin was suspended in DMF (20 mL) and shaken for 2 min. After filtering, 1:1 diethylamine:DMF (40 mL) was added to remove the Fmoc protecting groups and the resin was shaken for 1 hr. The resin was separated by filtration and washed with DMF (2x20 mL, 2 min each); 2:1 dioxane: water (2x20 mL, 5 min each), DMF (3x20 mL, 2 min each), DCM (3 x 20 mL, 2 min each) then dried in vacuo at 25° C. (The resin was found to have 0.4 mmol/g amino groups by picric acid titration at this stage.)

150 mg Portions of the resin were placed into seven Merrifield vessels and suspended in DCM (5 mL). The

appropriate identifiers were activated as their acyl carbonates as follows (for the first coupling): T1 (6.6 mg, 0.0098 mmol) was dissolved in anhydrous ether (2 mL) and pyridine (10 μL) was added. Isobutyl chloroformate (1.3 μL, 0.0096 mmol) was added as a solution in anhydrous ether (0.1 mL). The resulting mixture was stirred at 25° C for 1 hr. during which time a fine white precipitate formed. The stirring was stopped and the precipitate was allowed to settle for 30 min. Solutions of the acylcarbonates of T2 and T3 were prepared in the same way. Aliquots (0.25 mL) of the supernatant solution of activated identifiers were mixed to give the appropriate 3-bit binary tag codes and the appropriate coding mixtures of identifiers were added to each of the seven synthesis vessels. The vessels were shaken in the dark for 12 hrs, and then each was washed with DCM (4x10 mL, 2 min each). A solution of the symmetrical anhydride of an Nα-Fmoc amino acid in DCM (3 equivalents in 10 mL) was then added to the corresponding coded batch of resin and shaken for 20 min. 5% N,N-diisopropylethylamine in DCM (1 mL) was added and the mixture shaken until the resin gave a negative Kaiser test.

The resin batches were filtered and combined, and then washed with DCM (4x20 mL, 2 min each), isopropanol (2x20 mL, 2 min each), DCM (4x20 mL, 2 min each). The next cycle of labelling/coupling was initiated by Fmoc deprotection as described above.

After Fmoc deprotection of the residues in the last position of the peptide, the side chain functionality was deprotected by suspending the resin in DCM (10 mL), adding thioanisole (2 mL), ethanedithiol (0.5 mL) and trifluoroacetic acid (10 mL) then shaking for 1 hr at 25° C. The resin was then washed with DCM (6x20 mL, 2 min each) and dried.

Electron Capture Gas Chromatography Reading of Code

A single, selected bead was placed in a Pyrex capillary tube and washed with DMF (5x10 μL). The bead was then suspended in DMF (1 μL) and the capillary was sealed. The suspended bead was irradiated at 366 nm for 3 hrs to release the tag alcohols, and the capillary tube subsequently placed in a sand bath at 90° C for 2 hrs. The tube was opened and bis-trimethylsilyl acetamide (0.1 mL) was added to trimethylsilylate the tag alcohols. After centrifuging for 2 min., the tag solution above the bead (1 μL) was injected directly into an electron capture detection, capillary gas chromatograph for analysis. Gas chromatography was performed using a Hewlett Packard Series II Model 5890 gas chromatograph equipped with a 0.2 mmx20 m methylsilicone fused silica capillary column and an electron capture detector. Photolysis reactions were performed using a UVP "Black Ray" model UVL 56 hand-held 366 nm lamp.

Antibody Affinity Methods

The anti-C-myc peptide monoclonal antibody 9E10 was prepared from ascites fluid as described in Evans et al., Mol. Cell Biol., 5, 3610-3616 (1985) and Munro and Pelham, Cell, 48, 899-907 (1987). To test beads for binding to 9E10, beads were incubated in TBST [20 mM Tris-HCl (pH 7.5), 500 mM NaCl and 0.05% Tween-20] containing 1% bovine serum albumin (BSA) to block non-specific protein binding sites. The beads were then centrifuged, resuspended in a 1:200 dilution of 9E10 ascites fluid in TBST + 1% BSA and incubated overnight at 4°C. Beads were subsequently washed three times in TBST and incubated for 90 min. at room temperature in alkaline phosphatase-coupled goat antimouse IgG antibodies (Bio-Rad Laboratories), diluted 1:3000 in TBST + 1% BSA. After washing the beads twice in TBST and once in phosphatase buffer (100 mM Tris-HCl, pH 9.5, 100 mM NaCl and 5 mM $MgCl_2$), the beads were incubated 1 hr at room temperature in phosphatase buffer containing one one-hundreth part each of AP Color Reagents A & B (Bio-Rad Laboratories). To stop the reaction, the beads were washed twice in 20 mM sodium EDTA, pH 7.4. Solution phase affinities between 9E10 and various peptides were determined by a modification of the competitive ELISA assay described by Harlow et al., Antibodies: a Laboratory Manual, 570-573, Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

From a 30 mg sample of the combinatorial library of peptides, 40 individual beads were identified which stained on exposure to the anti-C-myc monoclonal antibody. Decoding of these positive-reacting beads established the ligand's reaction sequence as the myc epitope (EQKLISEEDL) or sequences that differed by one or two substituents among the three N-terminal residues.

EXAMPLE 6

23,540,625 Mixed Amide Library

The encoding technique was tested further by the preparation of a combinatorial library of 23,540,625 members consisting of peptides and other amide compounds.

The synthesis was carried out using 15 different reagents in 5 steps and 31 different reagents in the sixth step. Four identifiers were used to encode each of the 5 steps with 15 reagents and five identifiers were used in the final step with 31 reagents. A label set of 25 identifiers was therefore prepared. 2-Nitro-4-carboxybenzyl, O-aryl substituted ω-hydroxyalkyl carbonate identifiers were employed, where the tag components were comprised of an alkyl moiety of from 3 to 12 carbon atoms and the aryl moieties were (A) pentachlorophenyl, (B) 2,4,5-trichlorophenyl, (C) 2,4,6-trichlorophenyl, or (D) 2,6-dichloro-4-fluorophenyl. A set of 25 tags was prepared using appropriate alkyl chains lengths with A, B, C or D, separable using a 0.2 mMx25M methylsilicone GC

column. The chemical compositions of tags T1-T25 (where T1 represents the tag with the longest retention time, and T25 the tag with the shortest retention time) are summarized below:

T1	10A	T6	10C	T11	7B	T16	5C	T21	2B
T2	9A	T7	9B	T12	7C	T17	4B	T22	2C
T3	8A	T8	9C	T13	6B	T18	4C	T23	1B
T4	7A	T9	8B	T14	6C	T19	3B	T24	1C
T5	10B	T10	8C	T15	5B	T20	3C	T25	2D

The designations 10A, 9A, etc. are as described in Example 3.

The fifteen reagents used in the first five stages and the code identifying them are represented below where 1 represents the presence of tag and 0 the absence thereof.

REAGENT	CODE
L-serine	(0001)
D-serine	(0010)
L-glutamic acid	(0011)
D-glutamic acid	(0100)
L-glutamine	(0101)
D-glutamine	(0110)
L-lysine	(0111)
D-lysine	(1000)
L-Proline	(1001)
D-Proline	(1010)
L-phenylalanine	(1011)
D-phenylalanine	(1100)
3-amino-benzoic acid	(1101)
4-aminophenyl acetic acid	(1110)
3,5-diamino-benzoic acid	(1111)

The 31 reagents and the code representing them in the sixth stage are represented below:

REAGENT	CODE
L-serine	(00001)
D-serine	(00010)
L-glutamic acid	(00011)
D-glutamic acid	(00100)
L-glutamine	(00101)
D-glutamine	(00110)
L-lysine	(00111)
D-lysine	(01000)
L-proline	(01001)
D-proline	(01010)
L-phenylalanine	(01011)
D-phenylalanine	(01100)
3-amino-benzoic acid	(01101)
4-aminophenyl acetic acid	(01110)
3,5-diamino-benzoic acid	(01111)
Succinic Anhydride	(10000)
Tiglic acid	(10001)
2-pyrazine carboxylic acid	(10010)

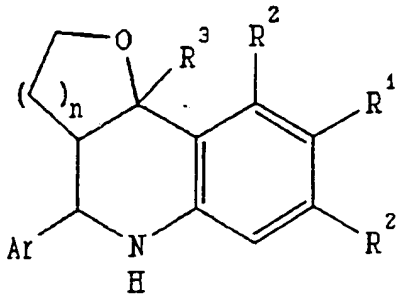
(±)thioctic acid	(10011)
1-piperidinepropionic acid	(10100)
piperonylic acid	(10101)
6-methylnicotinic acid	(10110)
3-(2-thienyl)acrylic acid	(10111)
methyl iodide	(11000)
tosyl chloride	(11001)
p-toluenesulfonyl isocyanate	(11010)
3-cyanobenzoic acid	(11011)
phthallic anhydride	(11100)
acetic anhydride	(11101)
ethyl chloroformate	(11110)
mesylchloride	(11111)

A spacer of six glycine units was prepared on the beads using standard methods. The variable region was constructed using butyl sidechain protection, and amino groups were protected as Fmoc derivatives. Amide bonds were formed by activation of the carboxylic acid with DIC and HOBt.

EXAMPLE 7

Hetero-Diels-Alder Library

A combinatorial hetero Diels-Alder library comprising 42 compounds of the formula:

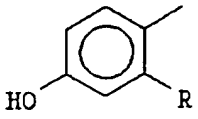
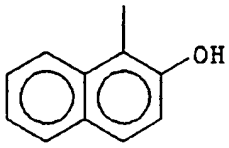


wherein;

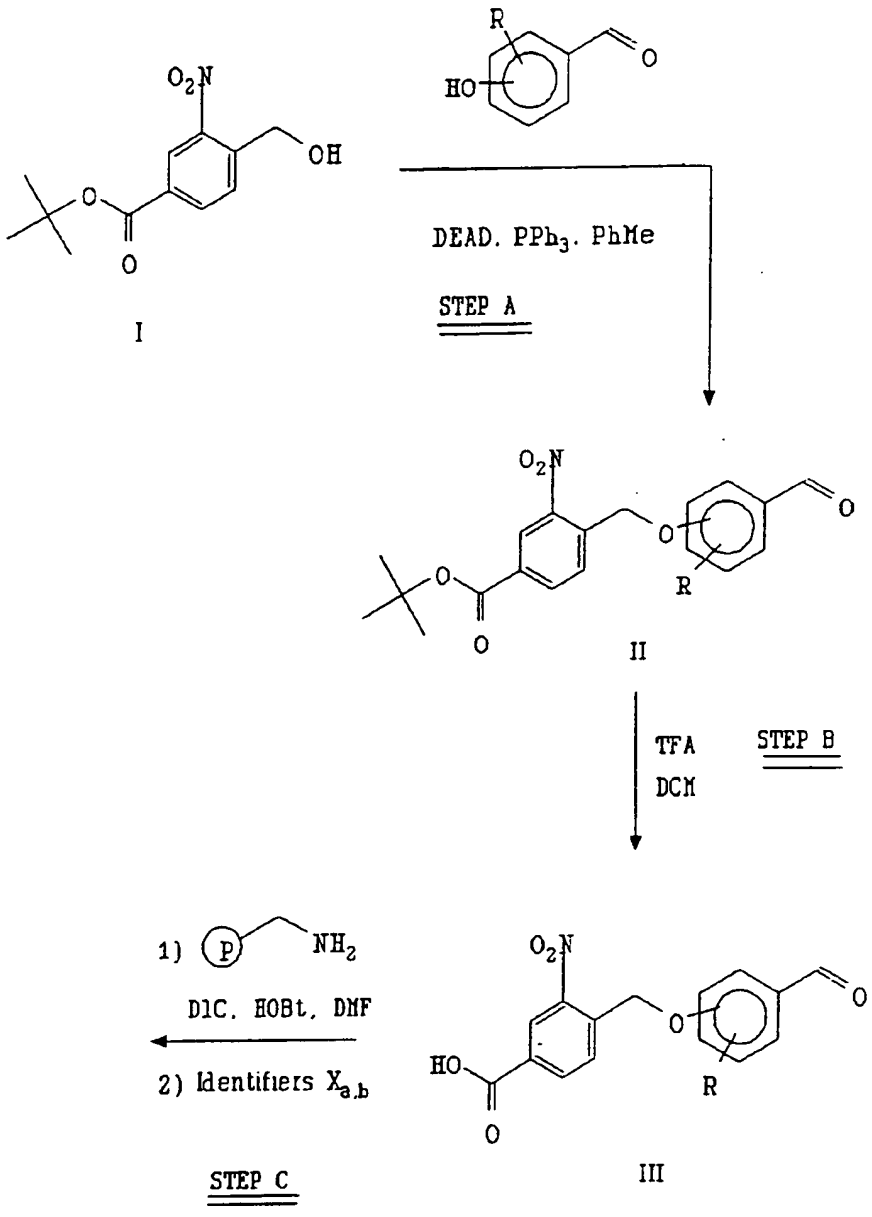
R^1 is H, CH_3O, F_3C, F_3CO, H_5C_6O, or C_6H_{11};

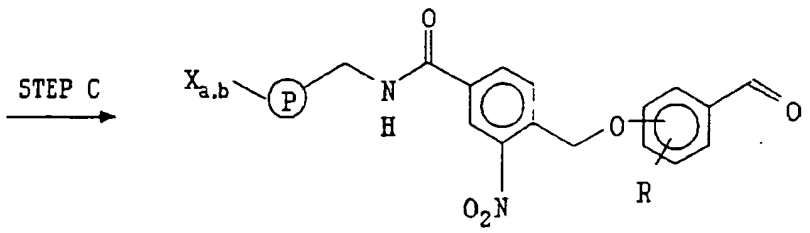
R^2 is H, CH_3, or CH_3O;

R^3 is H (when n=2), or CH_3 (when n=1); and

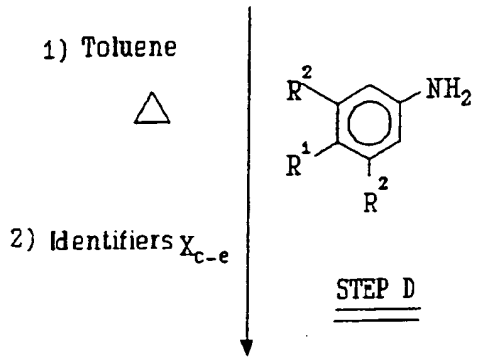
Ar =  R = H or Cl or Ar = 

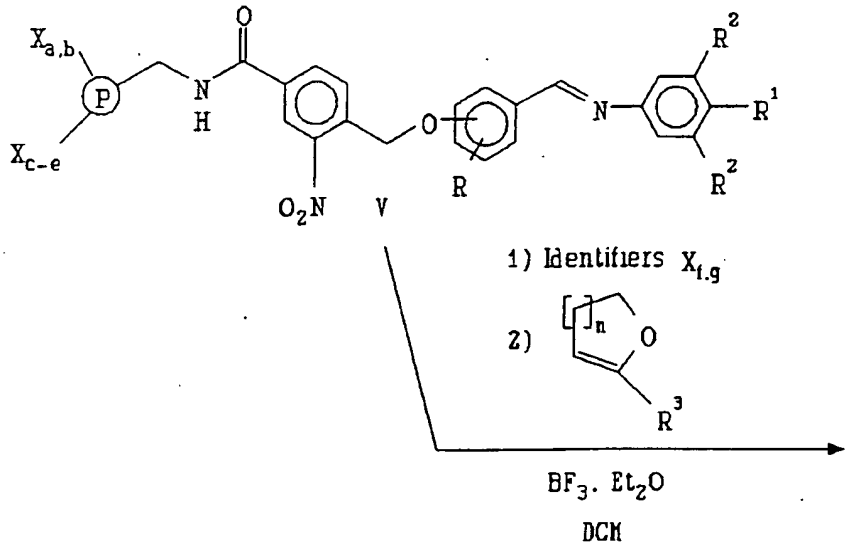
was constructed per the following scheme:



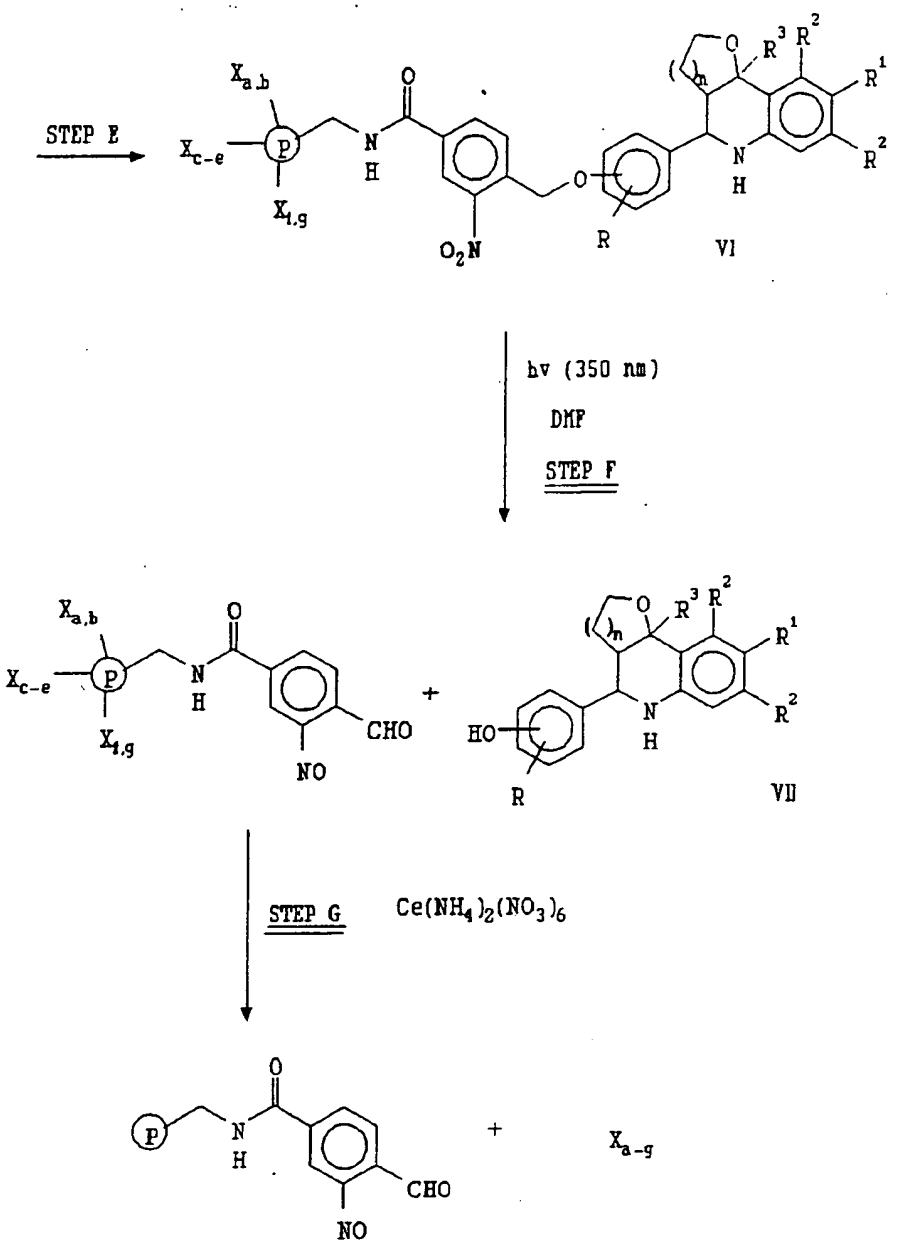

STEP C
$X_{a,b}$
P
N
H
O
O_2N
R
O

IV


1) Toluene
△
2) Identifiers X_{c-e}
R^2
NH_2
R^1
R^2
STEP D

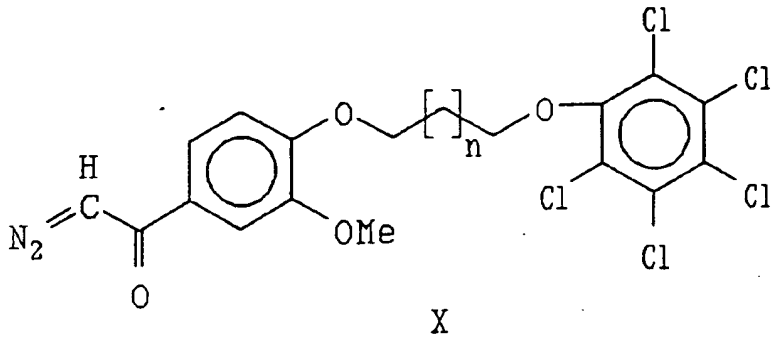

$X_{a,b}$
P
X_{c-e}
N
H
O
O_2N
V
O
R
N
R^2
R^1
R^2
1) Identifiers $X_{f,g}$
2)
n
O
R^3
$BF_3 . Et_2O$
DCM

STEP E


STEP E
hv (350 nm)
DMF
STEP F
STEP G
Ce(NH4)2(NO3)6
VI
VII
Xa-g

The azatricyclic products (VI) were constructed on polystyrene beads and were linked to the beads by a photocleavable linker allowing the azatricycle (VII) to be removed from the bead by exposure to U.V. light (350 nm in DMF). The binary codes introduced in steps C,D and E allow a unique determination of the reaction sequence used to introduce ArR, R^1, R^2 and R^3. The encoding tags were removed according to step G and analyzed by electron capture detection following GC separation.

The identifiers used in this scheme are represented by the formula X:



X

Wherein;

X_a indicates n=10

X_b indicates n=9

X_c indicates n=8

X_d indicates n=7

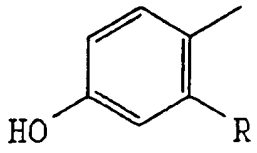
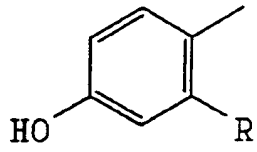
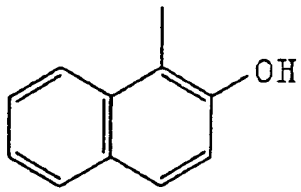
X_e indicates n=6

X_f indicates n=5

X_g indicates n=4

The codes for each of R, R^1, R^2, R^3 are as follows:

TABLE 7-1

X		
a	Ar = 	R = H
b	Ar = 	R = Cl
a,b	Ar = 	
c	R^1=H	R^2=H
d	R^1=H	R^2=CH_3
d,c	R^1=OCH_3	R^2=OCH_3

e	$R^1=CF_3$	$R^2=H$
e,c	$R^1=C_6H_5O$	$R^2=H$
e,d	$R^1=F_3CO$	$R^2=H$
e,d,c	$R^1=C_6H_{11}$	$R^2=H$
f	$R^3=CH_3$	n=1
g	$R^3=H$	n=2

Step A

To a solution of I (2.03 g, 8 mmol), 4-hydroxybenzaldehyde (1.17 g, 9.6 mmol) and triphenylphosphine (2.73 g, 10.4 mmol) in toluene (20 mL) stirring at 0°C was added over a period of 30 minutes diethylazodicarboxylate. The solution was allowed to warm and stirred for 1 hour once ambient temperature had been reached. The solution was concentrated by removal of approximately half of the solvent in vacuo and was then triturated with ether. The mixture was then filtered and the residue was washed thoroughly with ether. The solvent was removed in vacuo and the residue was purified by chromatography on silica gel (15% ethyl acetate in hexane) affording 1.3 g of the ether IIa (47% yield).

2-chloro-4-hydroxybenzaldehyde and 2-hydroxy-1-naphthaldehyde were coupled to I in analogous fashion affording ethers IIb and c in yields of 91% and 67%, respectively.

Step B

To a solution of ether IIa (0.407 g, 1.14 mmol) in DCM (20 mL) stirring at room temperature was added TFA (8 mL). The solution was allowed to stir for 6 hrs. The solution was evaporated to dryness in vacuo affording 0.343 g of acid IIIa (100% yield). Ethers IIb and IIc were deprotected analogously affording acids IIIb and c in yields of 92% and 100% respectively.

Step C

Into a peptide reaction vessel (Merrifield vessel) were measured 1% DVB (divinylbenzene) cross-linked polystyrene beads (50-80μ) functionalized with aminomethyl groups (1.1 meq/g) (200 mg of resin). The resin was suspended in DMF (2 mL) and shaken for 20 min. The acid IIIa (38 mg, 2 equiv.), 1-hydroxybenzotriazole (40 mg, 2 equiv) and diisopropylcarbodiimide (38 mg, 2 equiv) were added and the mixture shaken until a negative Ninhydrin test was achieved (22 hr). The solution was removed by filtration and the resin was washed with DCM (8x 10 mL).

The resin was resuspended in DCM (5 mL), identifier Xa (15 mg) was added and the flask was shaken for 1 hr. $Rh(TFA)_2$

catalyst (1 mol%) was added and the flasks shaken for 2 hrs. The solvent was removed by filtration and the resin resuspended in DCM (5 mL). Trifluoroacetic acid (1 drop) was added and the vessel shaken for 20 min. The solvent was removed by filtration, and the resin was washed with DCM (8x 10 mL).

In an analogous fashion, acids IIIb and IIIc were attached to the resin and were encoded with the appropriate identifiers, i.e., Xb for acid IIIb and Xa and Xb for acid IIIc. The three batches of resin were combined, mixed, washed, and dried.

Step D

The dry resin was divided into 7 equal portions (87 mg) which were put into seven peptide reaction vessels (Merrifield vessels) which were wrapped with heat tape. The resin in each vessel was suspended in toluene (10 mL) and shaken for 20 min. An appropriate amount of one aniline was then added to each flask (see Table 7-2).

TABLE 7-2

FLASK	ANILINE	AMOUNT ADDED
1	Aniline	3 mL
2	3,5-dimethylaniline	3 mL
3	3,4,5-trimethoxyaniline	2 g
4	4-trifluoromethylaniline	3 mL
5	4-phenoxyaniline	2 g
6	4-trifluoromethoxyaniline	3 mL
7	4-cyclohexylaniline	2 g

The heating tape was connected and the reaction mixtures shaken at 70°C for 18 hrs. The heat tape was disconnected and the solvent was removed by filtration and each batch of resin was washed with dry DCM (4x 10 mL), ether (10 mL), toluene (10 mL) and DCM (2x 10 mL). Each of the portions was then suspended in DCM (5 mL) and to each flask was added the appropriate identifier or combination of identifiers (Xc-e)(15 mg)(see Table 7-1). The flasks were shaken for 1 hr. and then $Rh(TFA)_2$ (1 mol%) was added to each flask and shaking continued for 2 hrs.

The solvent was then removed and each batch of resin was re-suspended in DCM (5 mL) to which was added TFA (1 drop). This mixture was shaken for 20 min., then the solvent was removed by filtration. The batches of resin were then washed (DCM, 1x 10 mL) and combined, washed again with DCM (3x 10 mL) and then dried thoroughly in vacuo.

Step E

The dried resin was divided into two equal portions (0.3 g) and each was placed in a peptide reaction vessel. The resin batches were washed with DCM (2x 10 mL) and then resuspended in DCM (5 mL). To one flask was added the identifier Xf (15 mg) and to the other was added Xg (15 mg). The flasks were shaken for 1 hr. prior to the

addition of $Rh(TFA)_2$ catalyst (1 mol%). The flasks were shaken for 2 hrs. and then the solvent was removed by filtration. Each batch of resin was washed with DCM (3x 10 mL), and each was then resuspended in DCM (5 mL).

The appropriate enol ether (1 mL)(see Table 7-1) was added to the flasks and the vessels shaken for 30 min. To each flask was added a solution of $BF_3 \cdot OEt_2$ (0.5 mL of a 5% solution in DCM) and the flasks were shaken for 24 hrs. Removal of the solvent by filtration was followed by washing of the resin with DCM (10 mL) and the resin was then combined. The beads were then washed further with DCM (5x 10 mL), DMF (2x 10 mL) methanol (2x 10 mL) and DCM (2x 10 mL). The resin was then dried thoroughly in vacuo.

Step F

To confirm the identity of the products produced in the Hetero-Diels-Alder library one example was completed on a large scale to allow confirmation of the structure by spectroscopic means. The procedure followed was essentially the same method as described for the combinatorial library. In step A 4-hydroxybenzaldehyde was coupled to the photolabile group. In step D, aniline was condensed with the aldehyde. In step E, the enol ether was formed with 4,5-dihydro-2-methylfuran.

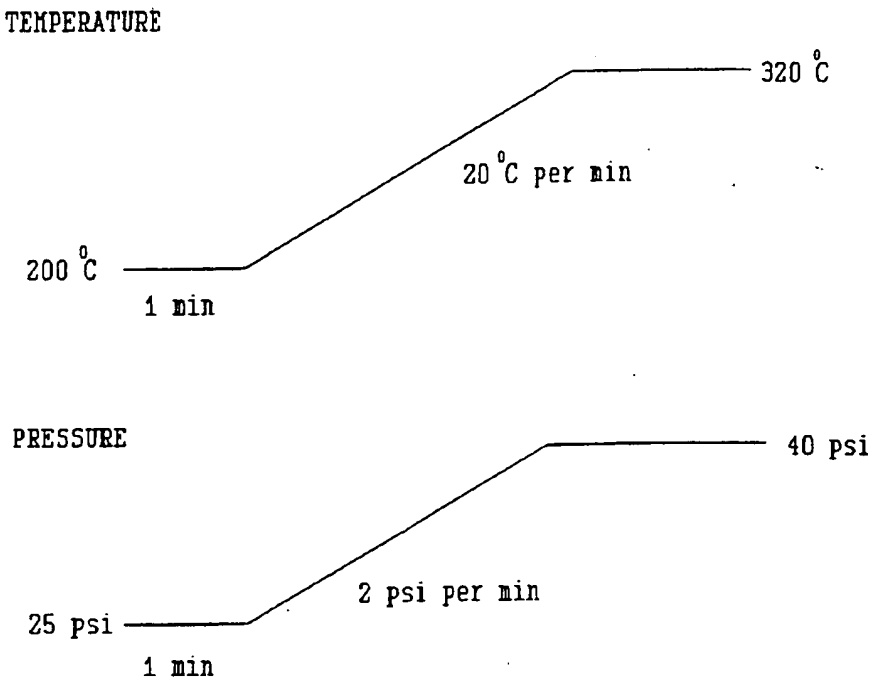
The photolysis of the compound (step F) was performed by suspending 100 mg of the beads in DMF (0.3 mL) and irradiating the beads with UVP "Black Ray" model UVL 56 hand-held 366 nm lamp for 16 hrs. The DMF was removed to one side by pipette and the beads rinsed with additional DMF (2x 3 mL). The original solution and the washings were combined and the solvent removed in vacuo. NMR analysis of the reaction mixture showed it to contain the desired azatricycle by comparison to the authentic sample.

Step G

A bead of interest was placed into a pyrex glass capillary tube sealed at one end. A solution (1 μL) of 1M aqueous cerium (IV) ammonium nitrate and acetonitrile (1:1) was syringed into the tube, and the tube was then centrifuged so that the bead lay on the bottom of the capillary and was completely immersed by the reagent solution. Hexane (2 μL) was added by syringe and the tube was again centrifuged. The open end of the capillary was flame-sealed and placed in an ultrasonic bath for 4 hrs. The capillary was then placed inverted into a centrifuge and spun such that the aqueous layer was forced through the hexane layer to the bottom of the tube. This extraction process was repeated 3 or 4 times and the tube was then opened. The hexane layer (1.5 μL) was removed by syringe and placed into a different capillary containing BSA (0.2 μL). This tube was sealed and centrifuged until the reagents were thoroughly mixed. A portion of the solution (ca. 1 μL) was removed and injected into a gas chromatography machine with a 25M x 0.2 mM methylsilicone fused silica column with electron capture detection for separation and interpretation of the tag molecules.

The sample was injected onto the GC column at 200°C and 25 psi of carrier gas (He_2). After 1 minute the temperature was increased at a rate of 20°C per minute to 320°C, and the pressure was increased at a rate of 2 psi per minute to 40 psi. These conditions are shown in the following diagram:

GC CONDITIONS



The following results were obtained with four randomly selected beads:

Bead 1

	TAG DETECTED		
	Xf	Xe Xd Xc	Xb Xa
Ar			2-Hydroxy naphthyl
R^1		C_6H_{11}	
R^2		H	
R^3	CH_3 (n=1)		

Bead 2

	TAG DETECTED		
	Xg	Xe Xd Xc	Xb
Ar			2-chloro-4-hydroxyphenyl
R^1		C_6H_{11}	
R^2		H	
R^3	H (n=2)		

Bead 3

	TAG DETECTED		
	Xg	Xe Xd	Xb Xa
Ar			2-Hydroxy naphthyl
R^1		F_3CO	
R^2		H	
R^3	H (n=2)		

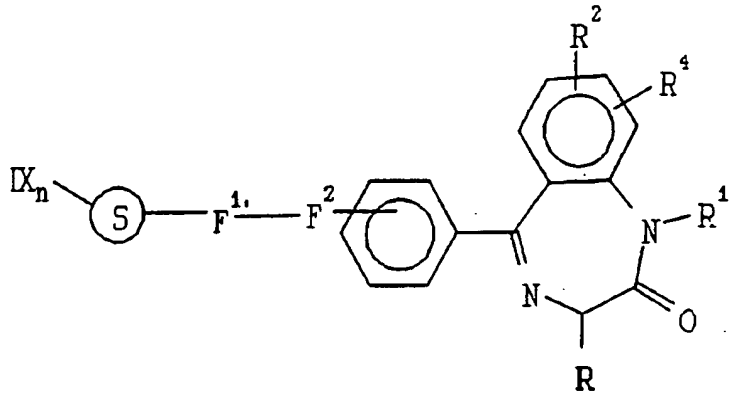
Bead 4

	TAG DETECTED		
	Xf	Xe Xd	Xb
Ar			2-chloro-4-hydroxyphenyl
R^1		F_3CO	
R^2		H	
R^3	CH_3 (n=1)		

EXAMPLE 8

Benzodiazepine Library

Following the procedure of Example 4, a combinatorial library is constructed of the Formula X



wherein

R is a radical of a naturally occurring D or L amino acid;

R^1 is H, C_1-C_6 alkyl, lower alkenyl, C_1-C_6 alkylamine, carboxy C_1-C_6 alkyl, or phenyl C_1-C_6 alkyl wherein the phenyl is optionally substituted by lower alkyl, F, Cl, Br, OH, NH_2, CO_2H, or O-lower alkyl;

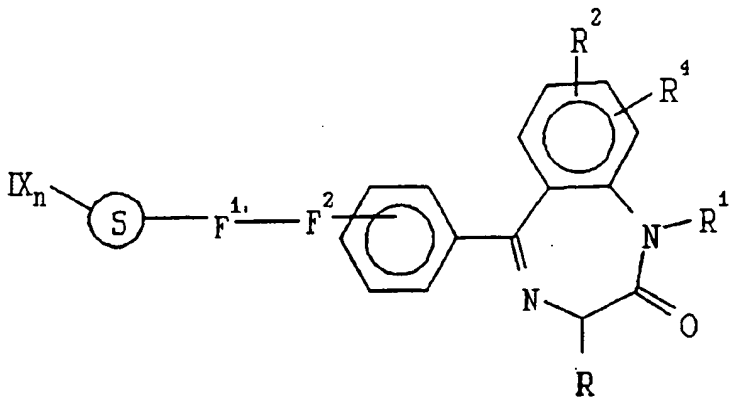
R^2 is H or CO_2H;

R^3 is H or OH;

R^4 is H or Cl;

with the provisos that when R^3 is OH, R^2 is H and when R^2 is carboxy, R^3 is H.

This library is released from a plurality of encoded beads of the general formula



wherein

IX_n is a plurality of identifiers of the Formula Ia wherein said plurality represents an encoded scheme;

S is a substrate;

$F^{1'}$-F^2 is the residue of the linker member of Formula Ia; and

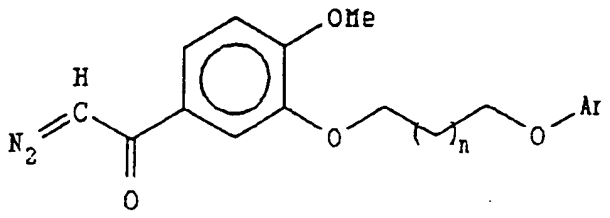
R, R^1, R^2, and R^4 are as defined for Formula X.

EXAMPLE 9

Typical Identifier Preparations

The diazo compound identifiers which are attached to the resin via carbene formation are prepared as exemplified.

Compounds of the general formula



wherein
n is 0-10 and
Ar is pentachlorophenol, 2,4,6-trichlorophenol, 2,4,5-trichlorophenol, or 2,6-dichloro-4-fluorophenol

are prepared as follows.

To a solution of 1-hydroxy-4-(2,6-dichloro-4-fluorophenoxy)butane (0.38 g, 1.5 mmol), methyl isovanillate (0.228 g, 1.5 mmol) and triphenylphosphine (0.393 g, 1.5 mmol) in THF (8 mL) was added diethylazodicarboxylate (0.287 g, 1.7 mmol). The solution stirred at r.t. for 36 hrs. The solvent was removed in vacuo and the residue purified by chromatography on silica gel (with a mixture of 20% ethyl acetate and 80% petroleum ether) affording 0.45 g of the aldehyde (77% yield).

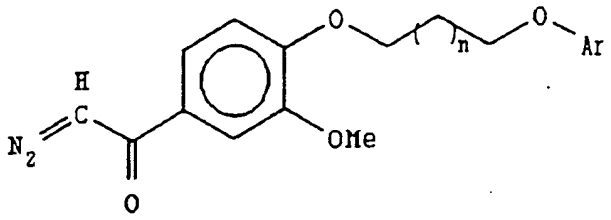
The aldehyde (100 mg, 0.26 mmol) was dissolved in acetone (8 mL) and was treated with a solution of $KMnO_4$ (61 mg, 0.39 mmol) in acetone (4 mL) and water (4 mL). The reaction stirred at room temperature for 13 hrs. The mixture was diluted with ethyl acetate (100 mL) and water

(50 mL) and the layers were separated. The aqueous layer was extracted with additional ethyl acetate (2x 100 mL). The combined organic layers were washed with water (50 mL) and dried ($MgSO_4$). Removal of the solvent afforded 109 mg of the benzoic acid (93% yield).

A solution of the acid (76 mg, 0.188 mmol) in methylene chloride (2 mL) was treated with oxalylchloride (36 mg, 0.28 mmol) and catalytic DMF. After stirring for 10 min at room temperature slow but steady evolution of gas was observed. Stirring continued for 2 hrs. when the solution was diluted with DCM (15 mL) and washed with saturated aqueous sodium hydrogencarbonate solution (5 mL). The layers were separated. The organic layer was dried (Na_2SO_4) and the solvent evaporated affording the benzoyl chloride as pale yellow crystals.

The benzoyl chloride was dissolved in methylene chloride (5 mL) and was added to a stirring solution of diazomethane in ether at -78°C. The cold bath was allowed to warm up and the mixture allowed to stir for 5 hrs at room temperature. The solvents and excess diazomethane were removed in vacuo and the residue purified by chromatography on silica gel using gradient elution method where the concentration of ethyl acetate ranged from 10% to 40 % in hexanes affording 48 mg of the diazo compound (60% yield).

Compounds of the general formula:



wherein;
n is 0-10 and
Ar is pentachlorophenol, 2,4,6-trichlorophenol, 2,4,5-trichlorophenol, or 2,6-dichloro-4-fluorophenol
are prepared as follows.

Methyl vanillate (0.729 g, 4.0 mmole), 1-hydroxy-9-(2,3,4,5,6-pentachlorophenoxy)nonane (1.634 g, 4.0 mmole) and triphenylphosphine (1.259 g, 4.8 mmole) were dissolved in 20 mL dry toluene under argon. DEAD (0.76 mL, 0.836 g, 4.8 mmole) was added dropwise, and the mixture was stirred at 25°C for one hour. The solution was concentrated to half volume and purified by flash chromatography eluting with DCM to give 1.0 g (1.7 mmole, 43%) of the product as a white crystalline solid.

The methyl ester above (1.0 g, 1.7 mmole) was dissolved in 50 mL THF, 2 mL water was added followed by lithium hydroxide (1.2 g, 50 mmole). The mixture was stirred at 25°C for one hour then refluxed for five hours. After cooling to 25°C the mixture was poured onto ethyl acetate (200 mL) and the solution was washed with 1 M HCl (50 mL x3) then sat. aq. NaCl (1x 50 mL) and dried over sodium

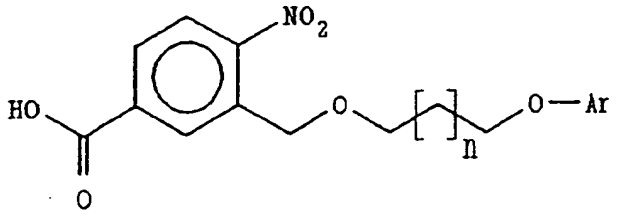
sulfate. The solvent was removed and the crude acid azeotroped once with toluene.

The crude material above was dissolved in 100 mL toluene, 10 mL (1.63 g, 14 mmole) thionyl chloride was added, and the mixture was refluxed for 90 min. The volume of the solution was reduced to approximately 30 mL by distillation, then the remaining toluene removed by evaporation. The crude acid chloride was dissolved in 20 mL dry DCM and cooled to -78°C under argon and a solution of approximately 10 mmole diazomethane in 50 mL anhydrous ether was added. The mixture was warmed to room temperature and stirred for 90 min. Argon was bubbled through the solution for 10 min. then the solvents were removed by evaporation and the crude material was purified by flash chromatography eluting with 10-20% ethyl acetate in hexane. The diazoketone (0.85 g, 1.4 mmole, 82% over three steps) was obtained as a pale yellow solid.

The following identifiers have been prepared as described above:

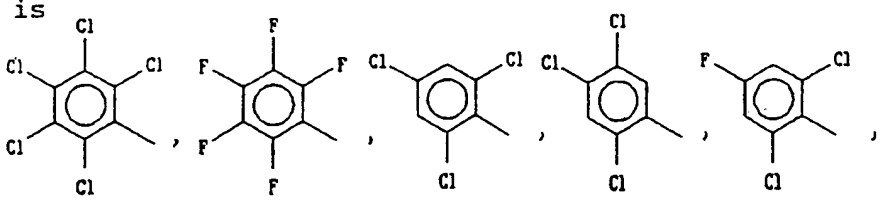
Photolabile Cleavage

50 Identifiers were prepared of the formula:



wherein:

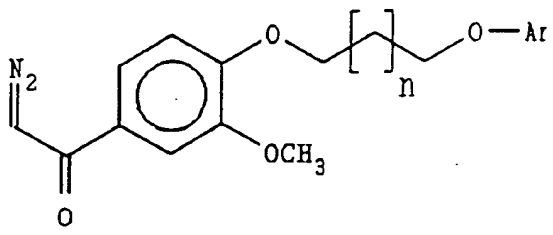
Ar is



and n is 1,2,3,4,5,6,7,8,9, and 10.

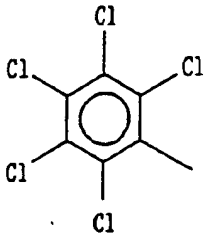
Oxidative Cleavage Type I

7 Identifiers were prepared of the formula



wherein:

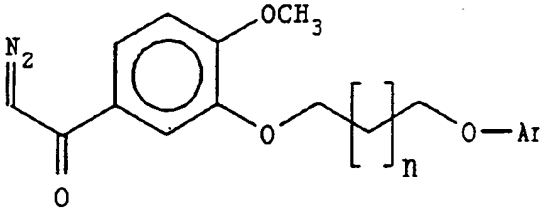
Ar is



and n is 4,5,6,7,8,9, and 10.

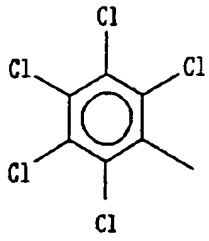
Oxidative Cleavage Type II

13 Identifiers were prepared of the formula



wherein:

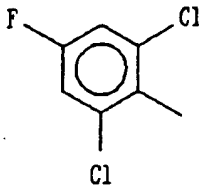
Ar is



and n is 1,2,3,4,5,6,7,8,9,10;

and wherein:

3Ar is



and n is 0,3, and 9.

It is evident from the above description that the subject invention provides a versatile, simple method for identifying compounds, where the amount of compound present precludes any assurance of the ability to obtain an accurate determination of its reaction history. The method allows for the production of extraordinarily large numbers of different products, which can be used in various screening techniques to determine biological or other activity of interest. The use of tags which are chemically inert under the process conditions allows for great versatility in a variety of environments produced by the various synthetic techniques employed for producing the products of interest. The tags can be readily synthesized and permit accurate analysis, so as to accurately define the nature of the composition.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

WHAT IS CLAIMED IS:

1. A method for recording the reaction history of a reaction series on each of a plurality of unique solid supports, wherein said reaction series involves at least two stages requiring differing agents or reaction conditions resulting in a different modification as to a plurality of said unique solid supports, resulting in a plurality of different final products on different unique solid supports, employing a combination of identifiers for recording said reaction history, said identifiers characterized by defining the choice of agent or reaction condition and the stage in said reaction series and being capable of being analyzed as to the choice and stage, said method comprising:

reacting, at a first or intermediate stage of said series, a different agent or employing a different reaction condition with each of a group of said unique solid supports, said group comprising at least one of said unique solid supports, and a combination of identifiers wherein said combination of identifiers defines the choice of agent and the stage in said series as to each group of said unique solid supports, each of said identifiers being individually bound to said unique solid support directly or through other than a prior identifier;

mixing said groups together and then dividing said plurality of unique solid supports into a plurality of groups for a second intermediate or final stage; and

repeating said reacting at least once to provide a plurality of final products, having different products on the different individual unique solid supports.

2. A method according to Claim 1, wherein at least 100 unique solid supports and at least 2 groups are employed in each said reacting.

3. A method according to Claim 1, including the additional stages of screening said final products on said unique solid supports for a characteristic of interest; and identifying the reaction history of at least one final product having said characteristic of interest.

4. A method of Claim 1 further comprising cleaving the final product from the solid support and screening said final product.

5. A method of Claim 1 further comprising treating the identifiers so as to detach the tag components from the solid supports and reacting said tag components with a moiety capable of detection by fluorescence or electron capture.

6. A method of Claim 5, wherein the detaching is done photochemically or oxidatively and the detectable moiety is derived from dansyl chloride or a polyhalobenzoylhalide.

7. A method according to Claim 5, wherein said tag components have two characteristics, a characteristic capable of separation and a characteristic capable of detection.

8. A method according to Claim 7, wherein said characteristic capable of detection is the ability to be detected by electron capture.

9. A method according to Claim 7, wherein said characteristic capable of detection is the ability to be detected by mass spectroscopy.

10. A method according to Claim 7, wherein said characteristic capable of detection is radioactivity.

11. A method according to Claim 7, wherein said characteristic capable of detection is fluorescence.

12. A method according to Claim 7, wherein said tags may be separated by means of chromatography.

13. A kit comprising a plurality of different separated organic compounds, each of the compounds characterized by having a distinguishable composition, encoding at least one bit of different information which can be determined by a physical measurement and sharing at least one common functionality.

14. A kit of Claim 13 comprising at least 4 different functional organic compounds.

15. A kit according to Claim 13, wherein said functional organic compounds are of the formula:

$$F^1-F^2-\underline{C}-E-\underline{C}'$$

where F^1-F^2 is a linker which allows for attachment to and detachment from a solid particle; and
$\underline{C}-E-\underline{C}'$ is a tag which can be determined by a physical measurement.

16. A kit according to Claim 15, wherein said functional organic compounds differ by the number of methylene groups and/or halogens, nitrogens or sulfurs present.

17. A kit according to Claim 15 wherein the C-E-C' portion can be removed photochemically.

18. A kit according to Claim 15 wherein the C-E-C' portion can be removed oxidatively, hydrolytically, thermolytically, or reductively.

19. A solid support characterized by having a ligand bound thereto and having a combination of identifiers bound to said solid support.

20. A solid support according to Claim 19, wherein said ligand is an oligomer which is an oligopeptide, oligonucleotide, oligosaccharide, polylipid, polyester, polyamide, polyurethane, polyurea, polyether, poly (phosphorus derivative) which is a phosphate, phosphonate, phosphoramide, phosphonamidey, phosphite, or phosphinamide, poly (sulfur derivative) which is a sulfone, sulfonate, sulfite, sulfonamide, or sulfenamide, where for the phosphorous and sulfur derivatives the indicated heteroatom for the most part will be bonded to C, H, N, O or S, and combinations thereof.

21. A solid support according to claim 19 wherein said ligand is a non-oligomer which is heterocyclic, aromatic, alicyclic, or aliphatic, and combinations thereof.

22. A solid support of Claim 21 wherein the non-oligomer is a diazabicyclic, an azatricyclic, or a branched amide compound.

23. A solid support of Claim 19 wherein the ligand is linked to the support through a non-labile linkage.

24. A solid support of Claim 19 wherein the ligand is linked to the support through a cleavable linkage.

25. A solid support according to Claim 19, wherein the identifier comprises tags, the tags being radioisotopes, or haloalkyl or haloarylallyl containing compounds.

26. A solid support of Claim 19 which is a bead of about 10-2000 μm in diameter, and wherein the identifiers comprise tag components which after cleavage from the bead can be separated by gas chromatography and or liquid chromatography detected by electron capture, mass spectroscopy, fluorescence, or atomic emission techniques.

27. A library comprising a plurality of solid supports according to claim 22.

28. A library of Claim 27, wherein the final products have been cleaved from the solid support.

29. A library of Claim 28, wherein the final products are a diazabicyclic, azatricyclic, or branched amide compounds.

30. A process for identifying compounds having a characteristic of interest which comprises screening a library of Claim 27.

31. A process of Claim 30, wherein the compounds have been cleaved from the solid surface.

32. A process of Claim 31, wherein the compound is a diazabicyclic, azatricyclic, or branched amide compound.

33. A method for producing a ligand involving a reaction series employing a method for recording the reaction history of a reaction series on each of a plurality of unique solid supports, wherein said reaction series involves at least two stages requiring differing agents and/or reaction conditions resulting in a different modification as to a plurality of said unique solid supports, resulting in a plurality of different final products on different unique solid supports, employing a combination of identifiers for recording said reaction history, said identifiers characterized by defining the choice of agent or reaction condition and the stage in said series and being capable of being analyzed as to the choice and stage, said method comprising:

reacting, at a first or intermediate stage of said series, a different agent or employing a different reaction condition with each of a group of said unique solid supports, said group comprising at least one of said unique solid supports, and a combination of identifiers wherein said combination of identifiers defines the choice of agent and the stage in said series as to each group of said unique solid supports, each of said identifiers being individually bound to said unique solid support directly or through other than a prior identifier;

mixing said groups together and then dividing said plurality of unique solid supports into a plurality of groups for a second intermediate or final stage;

repeating said reacting at least once to provide a plurality of ligands, having different products on the different individual unique solid surfaces; and

identifying said reaction history of at least one selected unique solid surface by means of said combination of identifiers.

34. A ligand according to Claim 33, wherein said identifying includes the stage of screening said ligands for a characteristic of interest.

35. A method for producing a ligand involving a reaction series employing a method for recording the reaction history of a reaction series on each of a plurality of unique solid surfaces, wherein said reaction series involves at least two stages requiring differing agents and/or reaction conditions resulting in a different modification as to each of a plurality of said unique solid surfaces, resulting in a plurality of different ligands on different unique solid surfaces, employing combinations of identifiers for recording said reaction history, said combination of identifiers characterized by defining the choice of agent and/or reaction condition and the stage in said series and being capable of being analyzed as to the choice and stage, said method comprising:

reacting, at a first or intermediate stage of said series, a different agent and/or employing a different reaction condition with each of a group of said unique solid surfaces, said group comprising at least one of said unique solid surfaces, and a combination of identifiers wherein said combination of identifiers defines the choice of agent and the stage in said series as to each group of said unique solid surfaces, each of said identifiers being individually bound to said unique solid surface through other than a prior identifier by a cleavable link;

mixing said groups together and then dividing said plurality of unique solid surfaces into a plurality of groups for a second intermediate or final stage;

repeating said reacting to provide a plurality of ligands having different ligands on the different individual unique solid surfaces;

screening the ligands from a plurality of each of said unique solid surfaces for a characteristic of interest; and

identifying said reaction history of at least one selected unique solid surface having ligand having said characteristic of interest by detaching the tag members from said unique solid surface and identifying said tag members by means of a differing characteristic.

36. A method according to Claim 35, wherein said tags differ in an homologous series and are detected by electron capture gas chromatography or mass spectroscopy.

37. A compound of the Formula I:

$$F^1-F^2-\underline{C}-E-\underline{C}' \qquad I$$

where F^1-F^2 is a linker which allows for attachment to and detachment from a support; and

$\underline{C}-E-\underline{C}'$ is the tag which is capable of analysis;

E is a tag component which allows for detection, or allows for detection and provides for separation as a result of variable substitution;

$\underline{C}$ and $\underline{C}'$ are tag components which allow for individual detection;

F^2 is a linking component capable of being selectively cleaved to release the tag components; and

F^1 is a functional group which allows ready attachment of the compound to a synthesis support.

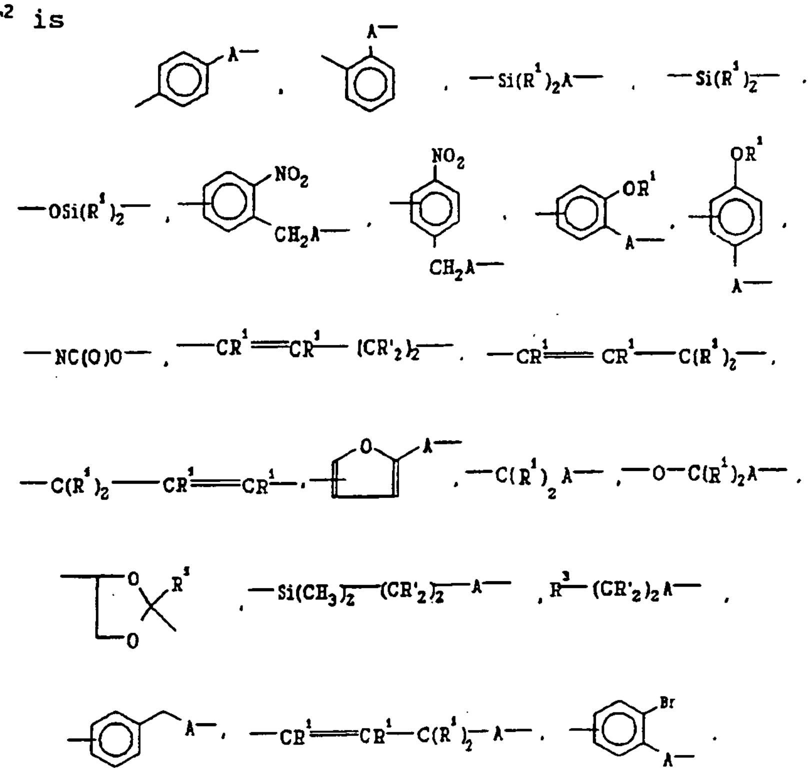
38. A compound of Claim 37 having the formula:

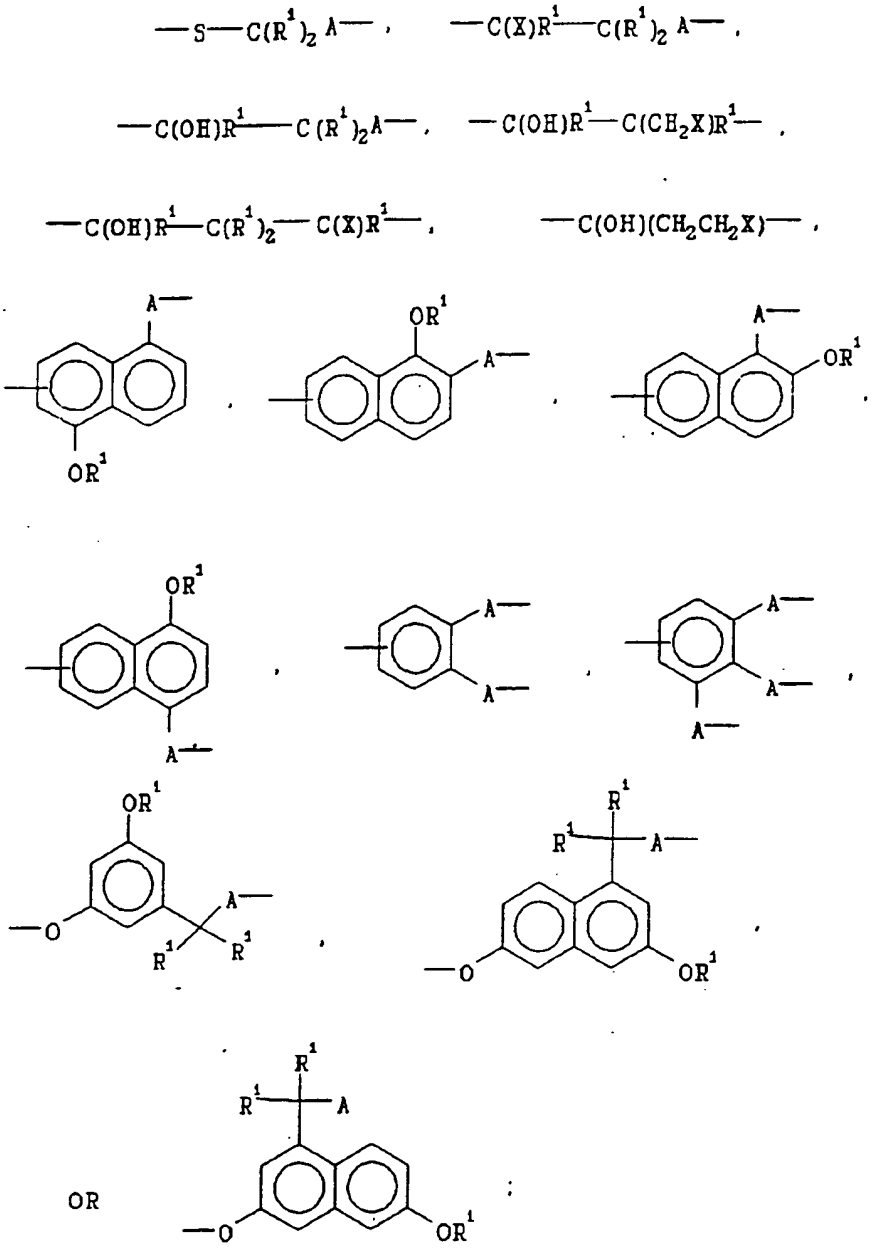
$$F^1-F^2-(\underline{C}(E-\underline{C}')_a)_b$$

wherein:

F^1 is CO_2H, CH_2X, NR^1R^1, $C(O)R^1$, OH, CHN_2, SH, $C(O)CHN_2$, $S(O_2)Cl$, $S(O_2)CHN_2$, N_3, NO_2, NO, $S(O_2)N_3$, OC(O)X, C(O)X, NCO, or NCS;

F^2 is





with the proviso that when F^2 is a bond, F^1 is OH or COOH;

A is -O, -OC(O)O-, -OC(O)-, or -NHC(O)-;

C is a bond, C_1-C_{20} alkylene optionally substituted by 1-40

F, Cl, Br, C_1-C_6 alkoxy, NR^4R^4, OR^4, or NR^4, or $-[(C(R^4)_2)_m-Y-Z-Y-(C(R^4)_2)_nY-Z-Y]_p-$; with the proviso that the maximum number of carbon atoms in C+C' is 20;

C' is H; F; Cl; C_1-C_{20} alkylene optionally substituted by

1-40 F, Cl, Br, C_1-C_6 alkoxy, NR^4R^4, OR^4, or NR^4, or $-[(C(R^4)_2)_m-Y-Z-Y-(C(R^4)_2)_nY-Z-Y]_p-$; with the proviso that the maximum number of carbon atoms in C+C' is 20;

E is C_1-C_{10} alkyl substituted by 1-20 F, Cl or Br; or Q-aryl

wherein the aryl is substituted by 1-7 F, Cl, NO_2, SO_2R^5, or substituted phenyl wherein the substituent is 1-5 F, Cl, NO_2, or SO_2R^5;

E-C' may be -H, -OH, or amino;

R^1 is H or C_1-C_6 alkyl;

R^3 is C=O, C(O)O, $C(O)NR^1$, S, SO, or SO_2;

R^4 is H or C_1-C_6 alkyl;

R^5 is C_1-C_6 alkyl;

a is 1-5;

b is 1-3;

m and n is each 0-20;

p is 1-7;

Q is a bond, O, S, NR^4, C=O, $-C(O)NR^5$, $-NR^5C(O)-$, -C(O)O-,

or -OC(O)-;

X is a leaving group such as Br, Cl, triflate, mesylate,

tosylate, or $OC(O)OR^5$;

Y is a bond, O, S, or NR^4;

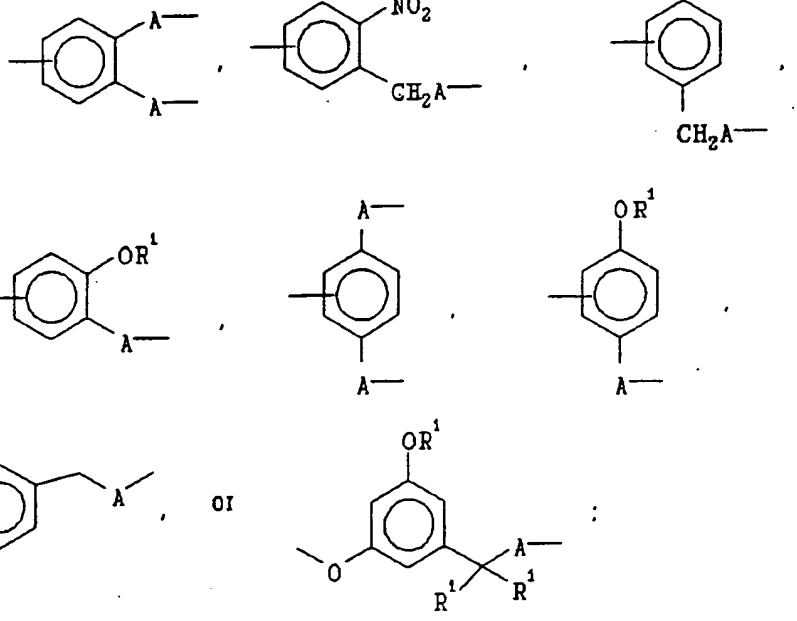
Z is a bond; phenyl optionally substituted by 1-4 F, Cl, Br, C_1-C_6 alkyl, C_1-C_6 alkoxy, C_1-C_6 alkyl substituted by 1-13 F, Cl, or C_1-C_6 alkyloxy substituted by 1-13 F, Cl, or Br; $(C(R^4)_2)_{1-20}$; or $(CF_2)_{1-20}$; with the proviso that when Z is a bond one of its adjacent Y's is also a bond and aryl is a mono- or bi-cyclic aromatic ring containing up to 10 carbon atoms and up to 2 heteroatoms selected from O, S, and N.

39. A compound of Claim 38 wherein:

F^1 is

CO_2H, OH, CHN_2, $C(O)CHN_2$, C(O)X, NCS, or CH_2X;

F^2 is

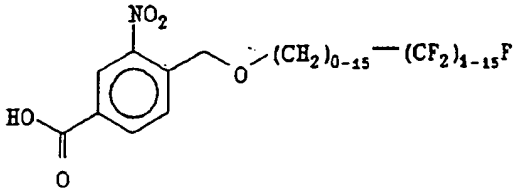


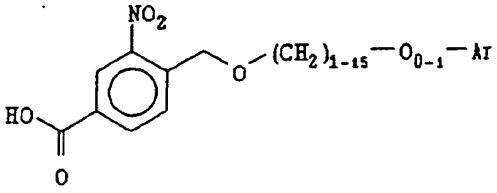
<u>C</u> and <u>C</u>' is each independently C_1-C_{20} alkylene unsubstituted or substituted by 1-40 F or Cl, or $[O-(CH_2)_{2-3}]_p$;

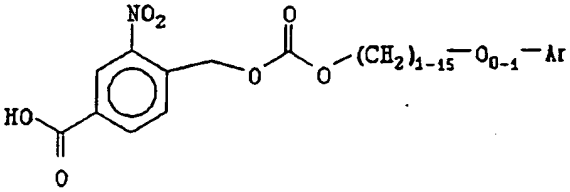
E is C_1-C_{10} alkyl substituted by 1-20 F or Cl; Q-aryl where aryl is a bi-cyclic aromatic ring substituted by 1-7 F or Cl; or Q-phenyl substituted by 1-5 F, Cl, NO_2, or SO_2R^5; and

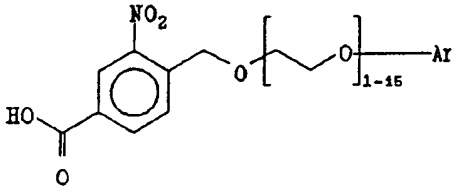
Q is a bond, O, $-NR^5C(O)-$, or $-OC(O)-$.

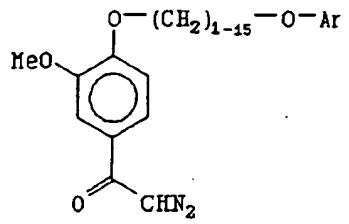
41. A compound of Claim 38 having the formula:

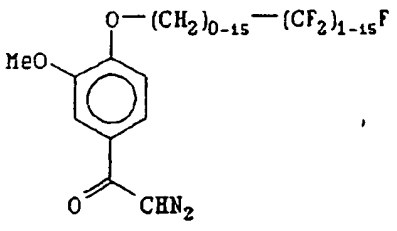




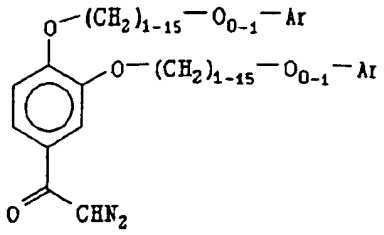








OR



wherein Ar is pentafluoro- pentachloro-, or pentabromophenyl, 2,3,5,6-tetrafluoro-4(2,3,4,5,6-pentafluorophenyl)phenyl, 2,4,6-trichlorophenyl, 2,4,5-trichlorophenyl, 2,6-dichloro-4-fluorophenyl, or 2,3,5,6-tetrafluorophenyl.

42. A compound of Claim 38 wherein:
E-C′ is H, OH, or NH_2.

43. A composition of the formula

$$S\text{-}F^{1\prime}\text{-}F^{2}\text{-}C\text{-}E\text{-}C'$$

wherein:

S is a soluble or solid support;

C-E-C′is the tag which is capable of analysis where E is a tag component which (a) allows for detection, such as an electrophoric group which can be analyzed by gas chromatography or mass spectroscopy or (b) allows for detection and for separation as a result of variable substitution;
C and C′ are tag components which allow for distinguishing one tag from all other tags, usually allowing for separation as a result of variable length or substitution, for example, varying the chromatographic retention time or the mass spectroscopy ratio Z/e;
F^2 is a linking component capable of being selectively cleaved to release the tag; and
$F^{1\prime}$ is a functional group which provides for attachment to the support.

44. A composition of claim 43 wherein:
S is a capillary, hollow fiber, needle, solid fiber, cellulose bead, pore-glass bead, silica gel, polystyrene bead optionally cross-linked with divinylbenzene, grafted co-poly bead, poly-acrylamide bead, latex bead, dimethylacrylamide bead optionally

cross-linked with N,N'-bis-acryloyl ethylene diamine, glass particles coated with a hydrophobic polymer, or low molecular weight non-cross-linked polystyrene; and

$P^{1\prime}$-F^2-C-E-C' is the residue of Formula I attached to S.

45. The method of Claim 1, wherein the combination of identifiers defines a binary coding scheme.

46. The method of Claim 1, wherein the identifiers are of Claim 37.

47. The method of Claim 1, wherein the identifiers are of Claim 38.

48. The method of Claim 1, wherein the identifiers are of Claim 39.

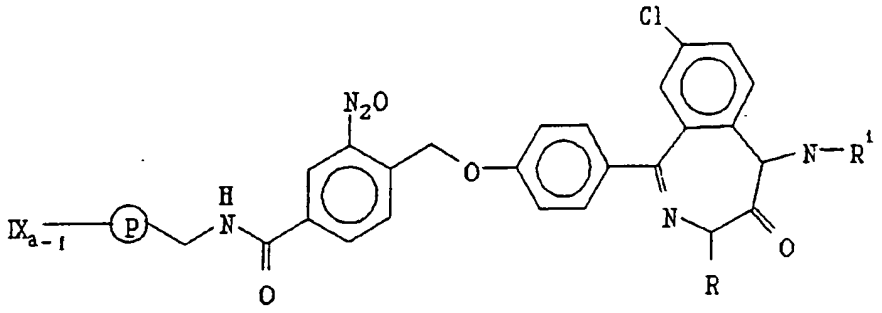
49. The method of Claim 1, wherein the identifiers are of Claim 42.

50. The method of Claim 1 further comprising detaching the tag members from said unique solid surfaces.

51. The method of Claim 50 wherein the tag members are detached photochemically, oxidatively, hydrolytically, thermolytically, or reductively.

52. The method of Claim 1 further comprising detaching non-oligomer ligands from said unique solid surfaces photochemically.

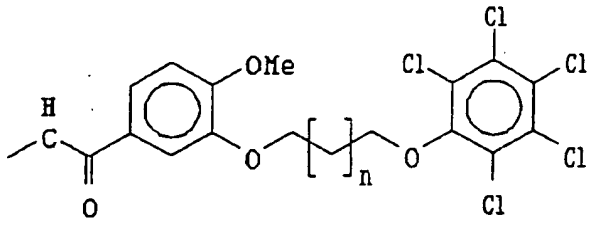
53. A compound of the formula



wherein:

P is a polystyrene resin;

IX_{a-f} is a plurality of residues of the formula



wherein:

n is 1 - 6;

R is CH_3, $CH(CH_3)_2$, CH_2CO_2H, $(CH_2)_4NH_2$, $CH_2-C_6H_4-OH$, or $CH_2C_6H_5$; and

R^1 is H, CH_3, C_2H_5, $CH_2CH=CH_2$, or $CH_2C_6H_5$.

54. A method of synthesizing a chemical compound so that the structure of the compound is readily determinable, which comprises synthesizing the compound on the surface of a solid support under conditions such that the solid support at the completion of the synthesis of the compound has bound to it a plurality of identifiers which encode the reaction stages associated with the synthesis of the compound.

55. A method of synthesizing a library of chemical compounds so that the structure of each compound in the library is readily determinable which comprises synthesizing each compound on the surface of a unique solid support under conditions such that each such unique support at the completion of the synthesis of the library of compounds has bound to it a plurality of identifiers which encode the reaction stages associated with the synthesis of the compound synthesized on such solid support.

56. A method of determining the structure of a chemical compound which comprises synthesizing the compound by the method of claim 54 or 55, isolating the solid support upon which the compound was synthesized, treating the solid support so isolated so as to cause the tag components of each of the identifiers bound to the solid support to be released, determining the identity or quantity or both of each tag component so released, and deriving the structure of the compound from the identities or quantities or both of all such tag components.

57. A method of identifying a compound having a desired characteristic which comprises synthesizing a library of chemical compounds by the method of claim 55, separately testing each of the compounds in the resulting library in an assay which identifies compounds having the desired characteristic so as to identify any compounds present in the library which has the desired characteristic.

58. A method of claim 57, further comprising determining the structure of the compound so identified.

59. A library of chemical compounds, each compound in the library being bound to a unique solid support and

each such solid support having bound to it a plurality of identifiers which encode the reaction stages associated with the synthesis of the compound bound to such solid support.

60. A library of claim 59, wherein compounds in the library are diazabicyclic compounds.

61. A library of claim 59, wherein compounds in the library are azatricyclic compounds.

62. A library of claim 59, wherein compounds in the library are branched amide compounds.

63. A library of claim 59, wherein compound in the library are peptides.

64. A method of identifying a compound having a desired characteristic which comprises testing a library of chemical compounds according to claim 58 in an assay which identifies compounds having the desired characteristic so as to identify any compound present in the library which have the desired characteristic.

65. A method of claim 64, further comprising determining the structure of the compound so identified.

66. A compound identified by the method of claim 63.

67. A method of claim 64, wherein the desired characteristic is antagonism for the human neurokinin 1/brandykin receptor and the library of chemical compounds comprises azatricyclic compounds.

68. A method of claim 64, wherein the desired characteristic in usefulness as a muscle relaxant, a

tranquilizer or a sedative and the library of chemical compounds comprising bezodiazopines.

69. A method of claim 64, wherein the desired characteristic is useful in the treatment of hypertension or Raynaud's syndrome and the library of chemical compounds comprises branched amides.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US93/09345

A. CLASSIFICATION OF SUBJECT MATTER
IPC(5) :C12Q 1/68; C07H 21/02, 21/04; A61K 48/00, 37/00
US CL :435/4, 6; 536/22.1; 514/2, 44
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/4, 6; 536/22.1; 514/2, 44

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS, BIOSIA, APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	P.N.A.S, VOLUME 89, ISSUED JUNE 1992, BRENNER ET AL, "ENCODED COMBINATORIAL CHEMISTRY", PAGES 5381-5383, SEE ENTIRE DOCUMENT.	1-41
Y	NATURE, VOLUME 354, ISSUED 07 NOVEMBER 1991, LAM ET AL., "A NEW TYPE OF SYNTHETIC PEPTIDE LIBRARY FOR IDENTIFYING LIGAND-BINDING ACTIVITY", PAGES 82-84, SEE PAGE 82, SUMMARY.	13-69.
Y	SCIENCE, VOLUME 257, ISSUED 17 JULY 1992, AMATO ET AL., "SPEEDING UP A CHEMICAL GAME OF CHANCE", PAGES 330-331, SEE ENTIRE DOCUMENT.	1-69.

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be part of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
20 December 1993

Date of mailing of the international search report
JAN 03 1994

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. NOT APPLICABLE

Authorized officer
EGGERTON CAMPBELL

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US93/09345

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	NATURE, VOLUME 354, ISSUED 07 NOVEMBER 1991, HOUGHTEN ET AL, "GENERATION AND USE OF SYNTHETIC PEPTIDE COMBINATORIAL LIBRARIES FOR BASIC RESEARCH AND DRUG DISCOVERY", PAGES 84-86, SEE ENTIRE DOCUMENT.	1-69.